الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
جامعة أبي بكر بلقايد–تلمسان
Université Aboubakr Belkaïd – Tlemcen –
Faculté de TECHNOLOGIE

MEMOIRE

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

En : Télécommunications

Spécialité : Réseaux et Télécommunications

Par Barka Mohammed Abdelillah & Boutlidja Abdelillah

Sujet

Détection des interférences dans un réseau 5G par Machine Learning

Soutenu publiquement, le 24/06/2026, devant le jury composé de :

A. BOUACHA	Professeur	Université de Tlemcen	Président
A. ABDELMALEK	MCB	Université de Tlemcen	Examinateur
F. DERRAZ	MCA	Université de Tlemcen	Encadrant
M. BOUSAHLA	MCB	Université de Tlemcen	Co-encadrant

Année universitaire : 2025–2026

Dédicace

Je dédie ce travail, avec tout mon respect et ma reconnaissance, à mes très chers parents, dont l'amour, les sacrifices et les encouragements ont accompagné chacune des étapes de mon parcours. Que ce mémoire soit le témoignage sincère de ma gratitude profonde et de ma volonté d'honorer leur confiance.

À mes frères et sœurs, pour leur affection, leur soutien moral et leur présence constante dans les moments importants.

À toute ma famille, qui m'a toujours encouragé à poursuivre mes études avec sérieux, patience et persévérance.

À mes amis et collègues, avec qui j'ai partagé des moments d'entraide, de travail et d'apprentissage.

À tous mes enseignants, qui ont contribué à ma formation scientifique et humaine.

Enfin, à toutes les personnes qui ont participé, de près ou de loin, à la réalisation de ce projet.

ABDELILLAH BOUTLIDJA

Dédicace

À mes très chers parents, pour leur amour inconditionnel, leur patience, leurs sacrifices et leur soutien indéfectible tout au long de mon parcours universitaire.

À mes frères et sœurs, dont la présence et les encouragements ont été une source permanente de motivation.

À mes enseignants et encadreurs, qui ont guidé mes pas avec rigueur, patience et bienveillance, et qui ont contribué à construire ma formation scientifique.

À tous mes amis et collègues, pour les moments partagés, l'esprit d'équipe et la solidarité tissée au fil des années.

Je dédie respectueusement ce modeste travail à toutes les personnes qui m'ont soutenu et encouragé jusqu'à l'aboutissement de ce mémoire.

Mohammed Abdelillah BARKA

Remerciements

Nous rendons grâce à Dieu Tout-Puissant pour nous avoir accordé la force, la patience et la persévérance nécessaires à la réalisation de ce mémoire.

Nous tenons à exprimer notre profonde gratitude à nos parents pour leur soutien indéfectible, leurs encouragements constants et leur patience tout au long de notre parcours universitaire.

Nous adressons nos sincères remerciements à nos encadreurs, Monsieur F. Derraz et Monsieur M. Bousahla, pour leur accompagnement et leurs précieux conseils.

Nous remercions également l'ensemble des enseignants de l'Université Abou Bekr Belkaid de Tlemcen, Faculté de Technologie, Département de Télécommunications, pour la qualité de leur enseignement et pour les connaissances qu'ils nous ont transmises durant nos années de formation.

Enfin, nous exprimons notre gratitude aux membres du jury pour l'honneur qu'ils nous font en évaluant ce mémoire, ainsi que pour leurs observations et recommandations qui contribueront à l'enrichissement de ce travail.

À toutes les personnes qui ont participé, de près ou de loin, à la réalisation de ce mémoire, nous adressons nos plus sincères remerciements.

Résumé

Ce mémoire traite de la détection automatique des interférences dans les réseaux 5G à partir de l'analyse des indicateurs radio de performance. Les variables exploitées incluent notamment le SINR/SINR,le RSSI, le RSRP, le RSRQ, le CQI et les débits montant et descendant. L'objectif principal est de construire une chaîne de traitement capable de distinguer un fonctionnement radio normal d'une situation perturbée, tout en conservant une interprétation compatible avec les grandeurs physiques utilisées en ingénierie des réseaux mobiles.

Le travail combine une analyse télécom des phénomènes d'interférence avec une méthodologie de Machine Learning complète. La chaîne proposée comprend le nettoyage des données, la construction des étiquettes, la normalisation, la sélection des variables, l'apprentissage supervisé, l'évaluation des performances et l'interprétation des résultats. Plusieurs modèles sont étudiés et comparés, notamment Random Forest, XGBoost et un réseau de neurones convolutif unidimensionnel (1D-CNN). Les performances sont analysées à l'aide de l'accuracy, de la précision, du rappel, du F1-score, des matrices de confusion et des courbes ROC.

Une attention particulière est accordée à l'explicabilité. Les méthodes d'importance des variables, SHAP et Grad-CAM sont utilisées afin d'identifier les KPI les plus influents et de relier les décisions des modèles à des indicateurs radio interprétables. Les résultats montrent que les approches de Machine Learning permettent de construire un système cohérent de détection des interférences 5G et de fournir une aide pertinente pour la supervision radio. Le mémoire souligne également les limites liées à la disponibilité des données, à la stratégie d'étiquetage et à la nécessité de validations complémentaires sur des mesures réelles diversifiées.

Mots-clés : réseaux 5G, interférences, Machine Learning, Random Forest, XGBoost, 1D-CNN, KPI radio, SHAP, Grad-CAM, explicabilité.

Abstract

This Master's thesis addresses the automatic detection of interference in 5G networks using radio Key Performance Indicators (KPIs). The considered variables include SINR/SINR, RSSI, RSRP, RSRQ, CQI, downlink throughput and uplink throughput. The main objective is to build a data-driven processing chain able to distinguish normal radio operation from degraded interference conditions while keeping the interpretation consistent with physical indicators used in mobile network engineering.

The proposed work combines a telecom-oriented analysis of interference phenomena with a complete Machine Learning methodology. The processing chain includes data cleaning, label construction, normalization, feature selection, supervised learning, performance evaluation and result interpretation. Several models are investigated and compared, including Random Forest, XGBoost and a one-dimensional Convolutional Neural Network (1D-CNN). Their performance is assessed using accuracy, precision, recall, F1-score, confusion matrices and ROC curves.

Particular attention is given to explainability. Feature importance, SHAP and Grad-CAM are used to identify the most influential KPIs and to relate model decisions to interpretable radio indicators. The obtained results show that Machine Learning approaches can support a coherent 5G interference detection system and provide useful assistance for radio network supervision. The thesis also highlights limitations related to data availability, labelling strategy and the need for additional validation using richer real-world measurements.

Keywords : 5G networks, interference, Machine Learning, Random Forest, XGBoost, 1D-CNN, radio KPIs, SHAP, Grad-CAM, explainability.

ملخص

يعالج هذا العمل موضوع الكشف الآلي عن التداخلات في شبكات الجيل الخامس بالاعتماد على مؤشرات الأداء الراديوية، مثل نسبة الإشارة إلى الضجيج، وقوة الإشارة المستقبلة، وجودة الإشارة، ومؤشر جودة القناة، ومعدل أخطاء الكتل، ومعدلات التدفق الصاعد والهابط. يهدف هذا البحث إلى بناء سلسلة معالجة قادرة على التمييز بين الحالة العادية للشبكة والحالة المتأثرة بالتداخلات، مع الحفاظ على تفسير فيزيائي واضح للنتائج انطلاقا من المؤشرات المستعملة في هندسة الشبكات الخلوية.

يجمع هذا العمل بين التحليل التقني لظواهر التداخل في شبكات الاتصالات ومنهجية كاملة قائمة على تعلم الآلة. وتشمل السلسلة المقترحة تنظيف البيانات، وبناء الوسوم، وتطبيع المتغيرات، واختيار الخصائص، وتدريب النماذج، وتقييم الأداء، ثم تفسير النتائج. وقد تمت دراسة ومقارنة عدة نماذج، من بينها الغابة العشوائية، ونموذج التعزيز الشجري، وشبكة عصبية التفافية أحادية البعد. وتم تقييم الأداء باستعمال مقاييس الدقة، والاسترجاع، ودقة التصنيف، ومؤشر التوازن بين الدقة والاسترجاع، ومصفوفات الالتباس، ومنحنيات الأداء.

تم إيلاء اهتمام خاص لقابلية تفسير القرارات، حيث استعملت أهمية الخصائص وتقنيات تفسير النماذج لتحديد المؤشرات الراديوية الأكثر تأثيرا وربط قرارات النماذج بمعان هندسية قابلة للفهم. وتبين النتائج أن تقنيات تعلم الآلة يمكن أن تساهم في بناء نظام متماسك للكشف عن التداخلات في شبكات الجيل الخامس، كما يمكن أن تساعد في تحسين مراقبة جودة الخدمة. كما يبرز العمل بعض الحدود المتعلقة بتوفر البيانات، وطريقة بناء الوسوم، والحاجة إلى التحقق باستعمال قياسات ميدانية أكثر تنوعا.

الكلمات المفتاحية شبكات الجيل الخامس، التداخلات، تعلم الآلة، الغابة العشوائية، الشبكات العصبية الالتفافية، مؤشرات الأداء الراديوية، قابلية التفسير.

Table des matières

Dédicace : ABDELILLAH BOUTLIDJA i

Dédicace : Mohammed Abdelillah BARKA ii

Remerciements iii

Résumé iv

Abstract v

Résumé arabe vi

Liste des abréviations xiv

Nomenclature xvii

Introduction générale 1

1 Concepts fondamentaux des réseaux 5G et des interférences 3

1.1 Introduction 4

1.2 Évolution des réseaux cellulaires 1G à 4G 5

1.3 Introduction sur 5G 7

1.4 Les applications de la 5G 8

1.5 Bande fréquence (sub-6ghz - mmwave) 9

1.6 Technologies clés de la 5G 10

1.7 Architecture du réseau 5G 12

1.8 Les interférences dans les réseaux cellulaires 16

1.8.1 Définition des interférences 16

1.8.2 Types d'interférences 17

1.9 Paramètres radio utilisés pour l'analyse 18

1.10 Effets des interférences sur les performances du réseau (QoS) 20

1.10.1 Dégradation du débit 20

1.10.2 Augmentation de la latence 21

1.10.3 Augmentation du taux d'erreur 21

2 Machine Learning appliqué aux réseaux 5G **22**
2.1 Introduction au Machine Learning dans les réseaux 5G . . . 23
2.2 Données télécoms et problématique data . . . 24
2.3 Pipeline data et apprentissage automatique . . . 25
2.3.1 Collecte et ingestion des données . . . 26
2.3.2 Prétraitement et nettoyage . . . 26
2.3.3 Feature engineering télécom . . . 27
2.3.4 Entraînement, validation et test . . . 28
2.3.5 Évaluation des performances . . . 29
2.4 Typologie des méthodes de Machine Learning . . . 29
2.4.1 Apprentissage supervisé . . . 30
2.4.2 Apprentissage non supervisé . . . 31
2.4.3 Apprentissage profond . . . 31
2.4.4 Apprentissage par renforcement . . . 32
2.5 Algorithmes de classification et d'analyse utilisés en télécom . . . 33
2.5.1 Régression logistique . . . 33
2.5.2 Random Forest . . . 33
2.5.3 XGBoost . . . 34
2.5.4 SVM . . . 34
2.5.5 K plus proches voisins . . . 35
2.5.6 CNN et LSTM pour séries temporelles de KPI . . . 35
2.6 Application à la détection des interférences 5G . . . 36
2.7 Machine Learning pour l'optimisation des réseaux 5G . . . 37
2.8 Architecture data, MLOps et boucle fermée . . . 37
2.9 Comparaison des méthodes ML pour les réseaux 5G . . . 39
2.10 Revue synthétique des travaux existants . . . 40
2.11 Discussion et Limites des ML dans la 5G . . . 41
2.12 Conclusion . . . 41

3 Dataset et prétraitement **43**
3.1 Description du problème étudié . . . 43
3.2 Présentation du dataset utilisé . . . 44
3.2.1 Origine et nature du dataset . . . 44
3.2.2 Description desvariables . . . 45
3.3 Génération du dataset . . . 46
3.3.1 Modélisation physique simplifiée . . . 47
3.3.2 Hypothèses retenues . . . 48
3.4 Prétraitement des données . . . 49
3.4.1 Nettoyage des données . . . 49
3.4.2 Normalisation des données . . . 49
3.4.3 Séparation des données d'entraînement et de test . . . 50
3.4.4 Sélection des variables (Feature Selection) . . . 51

3.5 Conclusion 51

4 Implémentation, workflows et résultats expérimentaux 53
4.1 Introduction 53
4.2 Objectifs expérimentaux du chapitre 54
4.3 Organisation générale du notebook 55
4.4 Environnement de développement 56
4.4.1 Environnement logiciel 56
4.4.2 Contraintes d'exécution 56
4.5 Importation des données et contrôle initial 57
4.5.1 Logique d'importation 57
4.5.2 Profil statistique initial 58
4.6 Description des KPI exploités 58
4.7 Étiquetage des données 59
4.7.1 Problématique de l'étiquetage 59
4.7.2 Étiquetage binaire 60
4.7.3 Étiquetage multi-classe exploratoire 62
4.8 Feature engineering et préparation des variables 63
4.8.1 Vecteur d'entrée 63
4.8.2 Pipeline de prétraitement 64
4.8.3 Partition train/test et SMOTE 64
4.9 Architecture expérimentale globale 65
4.9.1 Pipeline d'entraînement 65
4.9.2 Pseudo-code général 66
4.10 Modèle Random Forest 66
4.10.1 Principe 66
4.10.2 Paramètres utilisés 67
4.11 Modèle XGBoost 67
4.11.1 Principe 67
4.11.2 Paramètres utilisés 68
4.12 Modèle 1D-CNN 68
4.12.1 Justification du 1D-CNN 68
4.12.2 Architecture et entraînement 69
4.13 Vote majoritaire et décision finale 70
4.14 Métriques d'évaluation 71
4.14.1 Matrice de confusion 71
4.14.2 Formules des métriques 71
4.15 Résultats expérimentaux 72
4.15.1 Synthèse quantitative 72
4.15.2 Analyse des matrices de confusion 73
4.15.3 Courbes ROC 74
4.15.4 Relation entre SINR et probabilité d'interférence 75

4.16 Explicabilité par SHAP 76
4.16.1 Principe général 76
4.16.2 Importance globale des variables 76
4.16.3 Diagramme beeswarm 77
4.16.4 Explication locale et analyse d'erreur 78
4.17 Explicabilité par Grad-CAM 78
4.17.1 Principe appliqué au 1D-CNN 78
4.17.2 Grad-CAM global 79
4.17.3 Grad-CAM sur un nouveau cas 80
4.18 Comparaison croisée des importances 80
4.19 Dashboard et monitoring opérationnel 82
4.19.1 Passage du notebook au monitoring 82
4.19.2 Indicateurs à afficher dans un tableau de bord 82
4.20 Analyse critique des résultats 83
4.20.1 Pourquoi les modèles d'arbres dominent 83
4.20.2 Risque de réapprentissage de la règle d'étiquetage 83
4.20.3 Interprétation opérationnelle 84
4.21 Limites identifiées 84
4.22 Perspectives d'amélioration 85
4.22.1 Validation scientifique renforcée 85
4.22.2 Modèles temporels 86
4.22.3 Optimisation des seuils et calibration 86
4.23 Synthèse scientifique du chapitre 86
4.24 Conclusion 87

Conclusion générale et perspectives **88**

Bibliographie **90**

A Annexe code : implémentation Python du système de détection **93**
A.1 Import des bibliothèques principales 93
A.2 Chargement ou génération du dataset 94
A.3 Construction d'une étiquette binaire d'interférence 94
A.4 Prétraitement et séparation apprentissage-test 95
A.5 Entraînement de Random Forest et XGBoost 95
A.6 Architecture 1D-CNN 96
A.7 Évaluation et explicabilité 96
A.8 Remarque sur la reproductibilité 97

B Annexe données : dictionnaire des variables 5G **98**

Table des figures

1.1 Évolution des réseaux mobiles [2] 7
1.2 les cas d'utilisation du la 5G [4] 9
1.3 Répartition de la bande fréquence (sub-6ghz - mmwave) [4], [7] 10
1.4 Éléments fondamentaux de la technologie 5G [8]. 12
1.5 5G NSA, 5G SA [10]. 13
1.6 Architecture du réseau 5G [11] 14
1.7 Modes de déploiement autonome (a) et non-autonome (b) du réseau 5G [13] 16

2.1 Sources de données télécoms exploitables pour l'apprentissage automatique dans un réseau 5G. 25
2.2 Pipeline data complet pour passer des mesures 5G à une décision réseau assistée par Machine Learning. 26
2.3 Exemples de transformation des KPI 5G en caractéristiques utilisables par un modèle d'apprentissage. 28
2.4 Correspondance entre familles de méthodes Machine Learning, types de données télécoms et usages 5G. 30
2.5 Principe d'une classification supervisée d'états réseau à partir de KPI 5G. . 31
2.6 Choix pratique des méthodes Machine Learning selon les cas d'usage télécoms et la nature des données disponibles. 36
2.7 Architecture data et MLOps pour l'exploitation des modèles Machine Learning dans un contexte télécom. 38
2.8 Principe d'une boucle fermée pour la supervision et l'optimisation intelligente d'un réseau 5G. 38

4.1 Pipeline de détection des interférences 5G par Machine Learning 54
4.2 Chaîne expérimentale initiale générée à partir du notebook. 55
4.3 Workflow d'importation des données : lecture CSV ou génération synthétique. 57
4.4 Pipeline d'étiquetage binaire utilisé pour construire la cible supervisée. . . . 60
4.5 Distribution binaire du dataset après étiquetage affiné. 61
4.6 Visualisation de la règle d'étiquetage dans le plan SINR-RSRQ.. 61
4.7 Pipeline exploratoire d'étiquetage multi-classe. 62
4.8 Répartition exploratoire des types d'interférences. 62
4.9 Pipeline de prétraitement et de feature engineering. 64
4.10 Workflow d'entraînement des trois modèles. 65

4.11 Principe simplifié du Random Forest appliqué aux KPI radio. 67
4.12 Principe simplifié de XGBoost : correction séquentielle des erreurs. 68
4.13 Architecture 1D-CNN utilisée pour traiter le vecteur des KPI. 69
4.14 Courbes d'apprentissage du 1D-CNN sur 30 époques. 70
4.15 Pipeline de décision par vote majoritaire entre Random Forest, XGBoost et 1D-CNN. 70
4.16 Pipeline d'évaluation : prédictions, probabilités, matrices et métriques. . . . 72
4.17 Comparaison graphique des métriques de performance. 73
4.18 Matrices de confusion obtenues sur le jeu de test. 73
4.19 Courbes ROC des trois modèles générées dans le notebook. 74
4.20 Courbes ROC focalisées sur Random Forest et XGBoost. 75
4.21 Relation entre SINR et probabilité d'interférence prédite. 75
4.22 Workflow d'explicabilité combinant SHAP, importance native et Grad-CAM. 76
4.23 Importance globale des variables selon SHAP pour XGBoost. 77
4.24 Diagramme SHAP beeswarm pour le modèle XGBoost. 77
4.25 Exemple d'analyse locale issue du notebook. 78
4.26 Importance globale des KPI selon le Grad-CAM moyen du 1D-CNN. . . . 79
4.27 Explication Grad-CAM pour un nouveau point de mesure classé interférence. 80
4.28 Comparaison des importances normalisées : Random Forest, SHAP et Grad-CAM. 81
4.29 Pipeline proposé pour un monitoring temps réel des interférences. 82
4.30 Risque méthodologique lié à l'étiquetage rule-based et corrections possibles. 83
4.31 Feuille de route proposée pour renforcer la validation scientifique. 85

Liste des tableaux

1.1 Architecture fonctionnelle du réseau 5G [12]. 15

2.1 Comparaison des principales méthodes Machine Learning dans les applications télécoms 5G. 39
2.2 Synthèse de travaux et initiatives liés à l'AI/ML dans les réseaux 5G. . . . 40

3.1 Variables principales utilisées pour la détection d'interférences 45
3.2 Règles physiques utilisées pour l'étiquetage des classes d'interférences . . . 47
3.3 Répartition train/test du dataset 5G 50

4.1 Objectifs expérimentaux associés au chapitre 4. 55
4.2 Stack logiciel utilisé pour l'implémentation. 56
4.3 Contraintes pratiques observées et conséquences sur le travail. 57
4.4 Statistiques descriptives principales observées dans le notebook. 58
4.5 Interprétation des KPI utilisés dans le chapitre. 59
4.6 Règles exploratoires de classification multi-classe. 63
4.7 Variables finales utilisées par les modèles. 64
4.8 Organisation des données après partition et rééquilibrage. 65
4.9 Pseudo-code global du pipeline expérimental. 66
4.10 Hyperparamètres du Random Forest. 67
4.11 Hyperparamètres du modèle XGBoost. 68
4.12 Architecture du 1D-CNN. 69
4.13 Exemples d'interprétation du vote majoritaire. 71
4.14 Lecture de la matrice de confusion pour la détection d'interférence. 71
4.15 Performances des modèles sur le jeu de test. 72
4.16 Classement des KPI selon le Grad-CAM moyen du 1D-CNN. 79
4.17 Lecture comparative des méthodes d'explicabilité. 81
4.18 Indicateurs recommandés pour un tableau de bord opérationnel. 82
4.19 Limites principales du chapitre et pistes de correction. 84
4.20 Extensions possibles vers des modèles temporels. 86

B.1 Dictionnaire synthétique des variables radio utilisées. 98

Liste des abréviations

Abréviation	Signification
1D-CNN	One-Dimensional Convolutional Neural Network / réseau de neurones convolutionnel unidimensionnel
1G	First Generation / première génération des réseaux mobiles
2G	Second Generation / deuxième génération des réseaux mobiles
3G	Third Generation / troisième génération des réseaux mobiles
3GPP	Third Generation Partnership Project
4G	Fourth Generation / quatrième génération des réseaux mobiles
5G	Fifth Generation / cinquième génération des réseaux mobiles
5GC	5G Core / cœur de réseau 5G
AI	Artificial Intelligence / intelligence artificielle
AMF	Access and Mobility Management Function
AMPS	Advanced Mobile Phone System
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CAM	Class Activation Map
CDMA	Code Division Multiple Access
CNN	Convolutional Neural Network / réseau de neurones convolutionnel
CQI	Channel Quality Indicator / indicateur de qualité du canal
CPU	Central Processing Unit
CSV	Comma-Separated Values
DL	Downlink / lien descendant ; peut aussi désigner Deep Learning selon le contexte
EDGE	Enhanced Data rates for GSM Evolution
eMBB	enhanced Mobile Broadband
ETSI	European Telecommunications Standards Institute
F1-score	Score harmonique combinant précision et rappel
FDD	Frequency Division Duplex
FN	False Negative / faux négatif
FP	False Positive / faux positif
FR1	Frequency Range 1, bande 5G inférieure à 7,125 GHz
FR2	Frequency Range 2, bande 5G millimétrique
GPRS	General Packet Radio Service
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping

Abréviation	Signification
GRU	Gated Recurrent Unit
GSM	Global System for Mobile Communications
HSDPA	High Speed Downlink Packet Access
IA	Intelligence artificielle
IoT	Internet of Things / Internet des objets
KNN	k-Nearest Neighbors / méthode des k plus proches voisins
KPI	Key Performance Indicator / indicateur clé de performance
LTE	Long Term Evolution
LSTM	Long Short-Term Memory
MCS	Modulation and Coding Scheme
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning / apprentissage automatique
mmWave	Millimeter Wave / ondes millimétriques
mMTC	massive Machine-Type Communications
MMS	Multimedia Messaging Service
NG-RAN	Next Generation Radio Access Network
NR	New Radio
NSA	Non-Standalone
O-RAN	Open Radio Access Network
OFDM	Orthogonal Frequency Division Multiplexing
QoS	Quality of Service / qualité de service
RAN	Radio Access Network / réseau d'accès radio
RBF	Radial Basis Function
RB	Resource Block / bloc de ressources radio
RF	Radio Frequency / radiofréquence
RIC	RAN Intelligent Controller
RMS	Root Mean Square / valeur quadratique moyenne
ROC	Receiver Operating Characteristic
RRM	Radio Resource Management
RSNR	Reference Signal-to-Noise Ratio / rapport signal à bruit du signal de référence
RSRP	Reference Signal Received Power / puissance reçue du signal de référence
RSRQ	Reference Signal Received Quality / qualité reçue du signal de référence
RSSI	Received Signal Strength Indicator / indicateur de puissance totale reçue
SA	Standalone
SHAP	SHapley Additive exPlanations
SINR	Signal-to-Interference-plus-Noise Ratio / rapport signal sur interférence plus bruit
SMOTE	Synthetic Minority Over-sampling Technique
SMS	Short Message Service
SNR	Signal-to-Noise Ratio / rapport signal sur bruit
SON	Self-Organizing Network

Abréviation	Signification
SQL	Structured Query Language
SVM	Support Vector Machine / machine à vecteurs de support
TACS	Total Access Communication System
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TN	True Negative / vrai négatif
TP	True Positive / vrai positif
UE	User Equipment / équipement utilisateur
UIT	Union internationale des télécommunications
UL	Uplink / lien montant
UMTS	Universal Mobile Telecommunications System
uRLLC	Ultra-Reliable Low-Latency Communications
UPF	User Plane Function
XAI	Explainable Artificial Intelligence / intelligence artificielle explicable
XGBoost	Extreme Gradient Boosting
ZSM	Zero-touch network and Service Management

Nomenclature

Symbole / grandeur	Description	Unité
$\mathbf{x}_i$	Vecteur de caractéristiques associé à la i^{e} observation	–
x_{ij}	Valeur de la j^{e} variable pour la i^{e} observation	selon variable
z_{ij}	Valeur normalisée centrée-réduite de x_{ij}	–
y_i	Étiquette réelle de la i^{e} observation	–
$\hat{y}_i$	Classe prédite par le modèle	–
$\mathcal{D}$	Jeu de données constitué des couples $(\mathbf{x}_i, y_i)$	–
N	Nombre total d'observations du jeu de données	–
d	Nombre de variables ou de caractéristiques d'entrée	–
C	Nombre de classes dans un problème de classification multi-classe	–
K	Nombre de groupes dans un algorithme de clustering de type K-means	–
k	Nombre de voisins utilisés par l'algorithme KNN	–
M	Longueur d'une fenêtre temporelle ou nombre d'échantillons considérés	échantillons
n	Indice temporel discret	–
j	Indice de variable ou unité imaginaire selon le contexte mathématique	–
μ_j	Moyenne de la j^{e} variable calculée sur le jeu d'apprentissage	selon variable
σ_j	Écart-type de la j^{e} variable	selon variable
μ_{train}	Moyenne estimée sur l'ensemble d'apprentissage pour la normalisation	selon variable
σ_{train}	Écart-type estimé sur l'ensemble d'apprentissage pour la normalisation	selon variable
$f_{\boldsymbol{\theta}}$	Modèle paramétré utilisé pour associer des KPI à une décision	–
$\boldsymbol{\theta}$	Ensemble des paramètres d'un modèle d'apprentissage	–
$\boldsymbol{\theta}^\star$	Paramètres optimaux obtenus après entraînement	–
$\mathcal{L}$	Fonction de perte utilisée pendant l'apprentissage	–

Symbole / grandeur	Description	Unité
$\Omega(\boldsymbol{\theta})$	Terme de régularisation du modèle	–
λ	Coefficient de pondération de la régularisation	–
γ	Seuil, facteur d'actualisation ou paramètre de noyau selon le contexte	–
$\mathbf{W}^{(\ell)}$	Matrice des poids de la couche ℓ d'un réseau de neurones	–
$\mathbf{b}^{(\ell)}$	Vecteur de biais de la couche ℓ d'un réseau de neurones	–
ϕ	Fonction d'activation d'un réseau de neurones	–
$\pi(a\|s)$	Politique d'un agent en apprentissage par renforcement	–
s_t	État du système à l'instant t	–
a_t	Action choisie par l'agent à l'instant t	–
r_t	Récompense reçue à l'instant t	–
G_t	Retour cumulé à partir de l'instant t	–
SINR	Rapport signal sur interférence plus bruit	dB ou ratio
SINR	Rapport signal sur bruit	dB ou ratio
RSRP	Puissance reçue du signal de référence	dBm
RSRQ	Qualité reçue du signal de référence	dB
RSSI	Puissance totale reçue dans la bande considérée	dBm
CQI	Indicateur de qualité du canal radio	–
BLER	Taux d'erreur bloc	% ou ratio
R_i	Débit mesuré ou débit associé à l'observation i	bit/s, Mbit/s
R_u	Débit de l'utilisateur u	bit/s
R_{ref}	Débit de référence utilisé dans un indicateur de risque	bit/s
τ_i	Latence associée à l'observation i	s ou ms
L_i	Indicateur de charge du réseau ou de la cellule	% ou ratio
DL	Débit descendant ou lien descendant utilisé comme variable d'entrée	bit/s, Mbit/s
UL	Débit montant ou lien montant utilisé comme variable d'entrée	bit/s, Mbit/s
$\mathrm{DL_UL_Ratio}$	Rapport entre le débit descendant et le débit montant	–
$\mathrm{DL_bitrate}$	Débit binaire descendant utilisé dans le dataset	bit/s, Mbit/s
$\mathrm{UL_bitrate}$	Débit binaire montant utilisé dans le dataset	bit/s, Mbit/s
SINR_CQI_Product	Variable construite par produit entre le SINRet le CQI	–
$X[k]$	Transformée de Fourier discrète du signal $x[n]$	selon signal
$P[k]$	Puissance spectrale estimée au bin fréquentiel k	selon signal
$\mathcal{B}$	Bande fréquentielle étudiée	Hz
$\eta_{\mathcal{B}}$	Part d'énergie contenue dans la bande fréquentielle $\mathcal{B}$	–
P_s	Puissance du signal utile	W ou dBm

Symbole / grandeur	**Description**	**Unité**
P_I	Puissance d'interférence	W ou dBm
P_N	Puissance de bruit	W ou dBm
P_b	Puissance émise par la station de base b	W ou dBm
P_u	Puissance allouée à l'utilisateur u	W ou dBm
$P_{\max}$	Puissance maximale autorisée	W ou dBm
$G_{u,b}$	Gain de canal entre l'utilisateur u et la station de base b	– ou dB
N_{RB}	Nombre de blocs de ressources radio	–
N_0	Densité spectrale de bruit	W/Hz
B	Bande passante	Hz
B_u	Bande passante allouée à l'utilisateur u	Hz
B_{tot}	Bande passante totale disponible	Hz
$\mathcal{I}$	Indicateur global de risque d'interférence	–
$\alpha_1, \alpha_2, \alpha_3$	Poids associés aux composantes de l'indicateur de risque	–
$\mathrm{SINR}_{\mathrm{ref}}$	Valeur de référence du SINR	dB ou ratio
$\mathrm{BLER}_{\max}$	Valeur maximale ou seuil de référence du BLER	% ou ratio
TP	Nombre de vrais positifs	–
TN	Nombre de vrais négatifs	–
FP	Nombre de faux positifs	–
FN	Nombre de faux négatifs	–
$\mathrm{Accuracy}$	Taux global de bonnes prédictions	% ou ratio
$\mathrm{Precision}$	Proportion d'alertes positives réellement correctes	% ou ratio
Recall	Proportion d'interférences réellement détectées	% ou ratio
F_1	Moyenne harmonique de la précision et du rappel	% ou ratio
AUC	Aire sous la courbe ROC	–

Introduction générale

La cinquième génération des réseaux mobiles constitue aujourd'hui l'une des infrastructures essentielles de la transformation numérique. Elle ne se limite pas à une augmentation du débit par rapport aux générations précédentes : elle introduit une architecture radio et coeur plus flexible, des déploiements denses, des bandes de fréquences sub-6 GHz et millimétriques, des antennes massives, le beamforming, le découpage logique du réseau et une automatisation de plus en plus poussée [8, 9]. Ces évolutions permettent de supporter des usages variés tels que le très haut débit mobile, les objets connectés massifs, les applications industrielles à faible latence, les services critiques et les environnements intelligents.

Cependant, cette richesse technologique s'accompagne d'une complexité accrue. Dans un réseau 5G dense, la qualité perçue par l'utilisateur dépend fortement des conditions radio locales, de la charge du réseau, de la mobilité, de la planification fréquentielle, de la coexistence entre cellules et de la présence d'interférences. Ces interférences peuvent être co-canal, inter-cellulaires, adjacentes, externes ou liées à une mauvaise configuration du réseau. Elles se traduisent généralement par une dégradation d'indicateurs mesurables tels que le SINR, le RSRP, le RSRQ, le RSSI, le CQI, le BLER et le débit utilisateur [14, 17, 29]. Leur détection constitue donc une étape importante pour maintenir la qualité de service et améliorer l'exploitation opérationnelle des réseaux mobiles.

Les méthodes classiques de détection reposent souvent sur des seuils fixes ou sur des règles d'ingénierie. Ces méthodes présentent l'avantage d'être simples et interprétables, mais elles deviennent insuffisantes lorsque les environnements radio sont très variables. Une même valeur de KPI peut avoir une signification différente selon la bande de fréquence, la densité de cellules, le contexte de mobilité ou le niveau de trafic. Il devient alors nécessaire d'exploiter des approches capables d'apprendre des relations non linéaires entre plusieurs indicateurs et de produire une décision plus robuste.

Le Machine Learning offre dans ce contexte une réponse intéressante. Les modèles supervisés peuvent apprendre à distinguer des états normaux et des états perturbés à partir d'exemples mesurés ou simulés. Les modèles ensemblistes, comme Random Forest et XGBoost, sont particulièrement adaptés aux données tabulaires de type KPI, car ils combinent performance, robustesse et possibilités d'analyse de l'importance des variables [35, 36]. Les réseaux de neurones, quant à eux, permettent d'étudier des représentations plus flexibles et d'introduire

des outils d'explicabilité spécifiques, notamment Grad-CAM pour les architectures convolutionnelles [37, 45]. Les méthodes d'explicabilité comme SHAP complètent cette démarche en reliant les décisions des modèles aux variables qui les influencent le plus [40].

L'objectif de ce mémoire de Master est de proposer, mettre en oeuvre et analyser une chaîne de détection des interférences dans les réseaux 5G à partir de KPIs radio et de techniques de Machine Learning. Le travail ne vise pas seulement à obtenir de bonnes métriques de classification ; il cherche également à construire un pipeline cohérent, reproductible et interprétable. Cette exigence est importante, car un système utile pour l'ingénierie réseau doit pouvoir expliquer pourquoi une alerte est déclenchée, quelles variables y contribuent et quelles limites doivent être prises en compte avant une validation terrain.

Les contributions principales de ce mémoire sont les suivantes :

- présenter les concepts fondamentaux des réseaux 5G, leurs architectures et les mécanismes d'interférences qui affectent les performances radio ;
- analyser les KPIs radio pertinents pour la détection des dégradations, en particulier SINR, RSRP, RSRQ, RSSI, CQI et les indicateurs de débit ;
- construire une méthodologie de préparation des données incluant le nettoyage, l'étiquetage, la normalisation, le rééquilibrage de l'apprentissage et la séparation stricte entre apprentissage et test ;
- comparer plusieurs modèles de classification, notamment Random Forest, XGBoost et un réseau 1D-CNN, sur un protocole commun d'évaluation ;
- intégrer une analyse d'explicabilité fondée sur l'importance des variables, SHAP et Grad-CAM afin de vérifier la cohérence physique des décisions ;
- fournir une base reproductible comprenant les figures, les tableaux, le code et les références bibliographiques nécessaires à la compilation du mémoire.

Le mémoire est organisé en quatre chapitres. Le premier chapitre présente les notions de base des réseaux cellulaires et de la 5G, ainsi que les principaux types d'interférences et leurs effets sur les indicateurs radio. Le deuxième chapitre introduit les méthodes de Machine Learning appliquées aux réseaux 5G, en mettant l'accent sur la supervision, la classification, l'apprentissage profond et l'automatisation réseau. Le troisième chapitre décrit le jeu de données, les KPIs utilisés, les hypothèses d'étiquetage et la préparation expérimentale. Le quatrième chapitre présente l'implémentation, les modèles, les résultats, les métriques et l'analyse d'explicabilité. Enfin, la conclusion générale résume les résultats obtenus et propose des perspectives d'amélioration, notamment l'utilisation de données réelles multi-sites, la validation par experts terrain, l'étude temporelle des KPIs et l'intégration future dans une boucle de supervision intelligente.

Chapitre 1

Concepts fondamentaux des réseaux 5G et des interférences



Résumé :

Ce chapitre présente les concepts fondamentaux nécessaires à la compréhension des réseaux mobiles de cinquième génération et des phénomènes d'interférences qui affectent leurs performances. Il commence par rappeler l'évolution des réseaux cellulaires, depuis la première génération analogique jusqu'aux systèmes 4G, afin de mettre en évidence les progrès réalisés en matière de débit, de couverture, de mobilité et de qualité de service. Le chapitre introduit ensuite les principales caractéristiques de la 5G, notamment les communications à très haut débit, la faible latence, la connectivité massive des objets, ainsi que les cas d'usage eMBB, mMTC et uRLLC. Une attention particulière est accordée aux bandes de fréquences sub-6 GHz et millimétriques, aux technologies clés telles que le Massive MIMO, le beamforming, le network slicing, l'edge computing et les architectures 5G SA/NSA. Enfin, le chapitre analyse les interférences dans les réseaux cellulaires, leurs origines, leurs types et leurs effets sur les indicateurs radio tels que SINR, RSSI, RSRP, RSRQ, CQI et BLER. Cette analyse permet d'établir le lien entre la dégradation radio et les performances réseau, constituant ainsi la base scientifique du système de détection par Machine Learning développé dans les chapitres suivants.

1.1 Introduction

L'architecture cellulaire forme aujourd'hui la trame de presque toutes les communications radio mobiles : appels, échanges de données ou services multimédias passent, dans la grande majorité des cas, par ce type de réseau. Elle permet à des terminaux en mouvement de rester connectés sans dépendre d'un point fixe, quelle que soit leur localisation dans la zone couverte. Le principe fondamental consiste à découper l'espace en cellules, chacune équipée d'une station de base reliée au cœur de réseau, de manière à réemployer intelligemment les mêmes bandes de fréquences et à augmenter la capacité globale.

À partir des premiers systèmes analogiques de type 1G, pensés presque uniquement pour la voix, l'ingénierie des réseaux cellulaires a avancé par paliers successifs. Les générations 2G, 3G, 4G puis 5G ont apporté, à chaque étape, des gains marqués en débit, en efficacité spectrale et en stabilité de la connexion, tout en réduisant la latence pour les services sensibles au temps. On est ainsi passé de la simple téléphonie mobile à l'accès généralisé à l'internet, puis à la connexion massive d'objets et, désormais, à des applications temps réel comme la réalité augmentée ou certains scénarios industriels.

Pour atteindre ces objectifs, les réseaux cellulaires s'appuient sur un ensemble de notions structurantes. La gestion fine des ressources radio détermine la façon dont le réseau attribue fréquences, puissances et créneaux temporels aux différents utilisateurs. La prise en charge de la mobilité permet aux terminaux de changer de cellule sans interruption de service, ce qui exige des procédures de handover efficaces. Le contrôle des interférences entre cellules, lui, conditionne directement la qualité des liaisons, surtout dans les zones à forte densité d'utilisateurs où de nombreux équipements partagent le même spectre. Sans une coordination précise de ces mécanismes, les performances chutent très vite.

Avec la 5G, cette logique franchit un nouveau cap. Les réseaux intègrent des antennes MIMO massives et des techniques de beamforming afin de concentrer l'énergie radio vers des directions spécifiques, ce qui améliore le débit perçu par chaque utilisateur et limite les perturbations entre voisins. Ils exploitent aussi des algorithmes d'apprentissage automatique pour adapter, presque en temps réel, la configuration du réseau à la charge, à la demande de trafic et aux conditions radio locales.

J'examine donc les réseaux cellulaires comme un ensemble cohérent de choix d'architecture, de protocoles et de techniques de traitement du signal, qui visent à répondre à une demande toujours plus forte en connectivité fiable et rapide. Comprendre leurs principes de fonctionnement permet d'analyser, avec un minimum de recul.

Ce chapitre explique les concepts fondamentaux des réseaux 5G, ainsi que leurs caractéristiques clés, leur architecture et les technologies 5G. Il en introduit également les interférences au sein des réseaux sans fil, qui est l'un des facteurs clés ayant une incidence sur la qualité de service et les performances globales.

1.2 Évolution des réseaux cellulaires 1G à 4G

1G (Première génération) : La première génération de téléphonie mobile (désignée par 1G) fonctionnait selon un mode analogique et comprenait des dispositifs relativement encombrants. Les normes principalement concernées étaient les suivantes : AMPS (Advanced Mobile Phone System), introduit en 1976 aux États-Unis, représente la première norme de réseau cellulaire. Principalement déployé en Amérique du Nord, en Russie et en Asie, ce réseau analogique de première génération présentait des mécanismes de sécurité limités, facilitant ainsi le piratage des lignes téléphoniques. Le système TACS (Total Access Communication System) constitue la déclinaison européenne du modèle AMPS. Employant la bande de fréquence de 900 MHz, ce système a été notamment largement déployé en Angleterre, puis en Asie (Hong Kong et Japon). ETACS (Extended Total Access Communication System) constitue une version perfectionnée de la norme TACS, élaborée au Royaume-Uni, qui exploite un nombre accru de canaux de communication. Les réseaux cellulaires de première génération ont été supplantés par l'émergence d'une seconde génération entièrement numérique [1].

2G (La deuxième génération de réseaux mobiles) : La deuxième génération de réseaux mobiles (désignée 2G) a constitué une rupture majeure par rapport à la première génération de téléphones cellulaires, en assurant la transition de la technologie analogique vers le numérique. Les normes fondamentales de la téléphonie mobile de deuxième génération (2G) sont les suivantes : Le GSM (Global System for Mobile Communications), norme prédominante en Europe à la fin du XXe siècle, est également pris en charge aux États-Unis. Ce standard exploite les bandes de fréquences de 900 MHz et 1800 MHz sur le territoire européen. Aux États-Unis, en revanche, les bandes de fréquences exploitées correspondent aux bandes de 850 MHz et 1900 MHz. Ainsi, les téléphones mobiles capables d'opérer à la fois en Europe et aux États-Unis sont désignés sous le terme de tri-bande. La technique CDMA (Code Division Multiple Access) repose sur un procédé d'étalement du spectre qui permet la transmission d'un signal radio sur une vaste bande de fréquences. Le TDMA (Accès Multiple par Répartition dans le Temps) repose sur une méthode de segmentation temporelle des canaux de communication afin d'accroître la quantité de données transmises simultanément. La technologie TDMA est essentiellement déployée sur le continent américain, ainsi qu'en Nouvelle-Zélande et dans la région Asie-Pacifique. Les réseaux 2G permettent la transmission de la voix ainsi que de données numériques de faible volume, telles que les messages texte (SMS, pour Short Message Service) et les messages multimédias (MMS, pour Multimedia Message Service). La norme GSM autorise un débit maximal de 9,6 kbps. Des prolongements de la norme GSM ont été développés dans le but d'accroître son débit. Cela concerne en particulier la norme GPRS (General Packet Radio System), qui autorise des débits théoriques d'environ 114 kbit/s, bien que la vitesse effective soit généralement plus proche de 40 kbit/s. Cette technologie, ne relevant pas de la classification « 3G », a été désignée sous le terme de 2,5G. La norme EDGE (Enhanced Data Rates for Global Evolution), désignée comme 2,75G, multiplie par quatre

les performances en débit de la norme GPRS en proposant un débit théorique de 384 Kbps, ce qui permet l'émergence d'applications multimédias. En réalité, la norme EDGE autorise des débits théoriques maximaux de 473 kbit/s, mais ceux-ci ont été restreints afin de satisfaire aux exigences des spécifications IMT-2000 (International Mobile Telecommunications-2000) de l'Union internationale des télécommunications (UIT) [1].

Troisième génération (3G) : Les spécifications IMT-2000 (International Mobile Télécommunications pour l'an 2000) de l'Union Internationale des Télécommunications (UIT) établissent les paramètres techniques de la troisième génération (3G) de téléphonie mobile. Ces attributs se distinguent notamment par les éléments suivants : Un débit de transmission élevé 144 Kbps offrant une couverture complète pour un usage mobile, 384 Kbps avec une couverture intermédiaire adaptée à une utilisation piétonne, et 2 Mbps avec une zone de couverture restreinte pour un usage fixe. Interopérabilité mondiale, compatibilité des services mobiles de troisième génération avec les réseaux de deuxième génération. La 3G permet d'atteindre des débits excédant 144 kbit/s, ce qui facilite l'émergence d'applications multimédias telles que la transmission vidéo, la visioconférence ou l'accès à Internet à haut débit. Les réseaux 3G exploitent des bandes de fréquences distinctes de celles des générations antérieures, comprises entre 1885-2025 MHz et 2110-2200 MHz. La norme 3G prédominante en Europe est désignée sous le nom d'UMTS (Universal Mobile Télécommunications System) et repose sur un codage W-CDMA (Wideband Code Division Multiple Access). La technologie UMTS exploite une bande de fréquence de 5 MHz pour la transmission de la voix et des données, avec des débits atteignant de 384 kbps à 2 Mbps. La technologie HSDPA

(High-Speed Downlink Packet Access), un protocole de téléphonie mobile de troisième génération désigné sous le terme « 3,5G », permet d'atteindre des débits situés entre 8 et 10 Mbit/s. La technologie HSDPA exploite la bande de fréquence de 5 GHz et applique le codage W-CDMA [1].

Quatrième génération de technologie mobile (4G) : La technologie 4G constitue la nouvelle génération des normes téléphoniques, actuellement en phase d'expansion à l'échelle mondiale, notamment en France où son accès demeure, pour l'instant, limité à quelques grandes agglomérations. Alors que, dans le réseau 3G, les données internet et les communications téléphoniques étaient distinctes, impliquant un partage du débit entre les utilisateurs connectés, la 4G intègre l'ensemble de ces données. Cela assure un transfert de données d'une qualité supérieure. Les débits varient de cent mégabits par seconde à un gigaoctet par seconde. Le réseau mobile 4G est ainsi en mesure de fournir un débit théoriquement supérieur à celui de la fibre optique actuelle, bien que, en pratique, il se limite pour l'instant à quelques dizaines de Mb/s en raison du partage de la bande passante entre les différents utilisateurs au sein d'une même zone. En France, la technologie 4G déployée correspond à la norme LTE (Long Term Evolution), laquelle exploite les bandes de fréquences de 2 600 MHz et de 800 MHz. À l'instar de la transition de la 2G à la 3G, les terminaux mobiles doivent être compatibles avec

la nouvelle génération 4G, condition remplie par un grand nombre de dispositifs récemment commercialisés. En effet, que ce soit une clé mobile ou un smartphone, leur compatibilité avec les nouveaux protocoles IPv6, rendue possible par la connexion 4G, s'avère indispensable. De leur côté, les opérateurs initient progressivement la mise en place de forfaits adaptés [1].

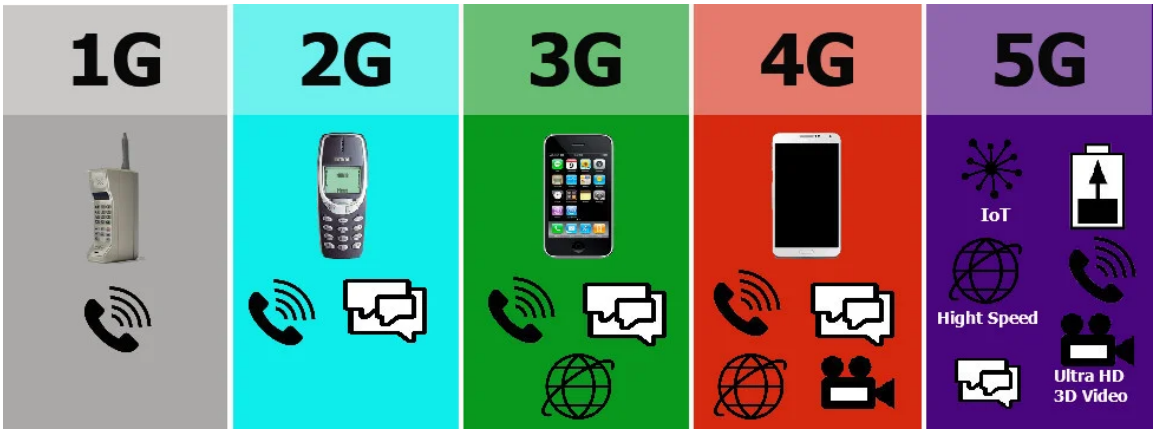


Fig. 1.1 – Évolution des réseaux mobiles [2]

1.3 Introduction sur 5G

La 5G représente un nouveau pas dans le développement de la télécommunication sans fil. Conçue pour résorber le phénomène de la progression continue de la connectivité, elle se distingue par des taux de transmission de données très élevés, des temps de latence inférieurs à quelques millisecondes, ainsi que par la capacité à gérer simultanément un nombre massif de terminaux. Par rapport à ses prédécesseurs, la 5G ne se contente plus de faciliter le développement des échanges entre les utilisateurs, mais vise à devenir le socle technique de base de l'Internet des objets (IoT), de l'environnement intelligent de ville, des transports intelligents, ainsi que de nombreuses

Formes de soins à distance. Autant de nouveaux services, somme toute logiques.

Avec le regard de la technologie, la 5G repose sur plusieurs briques clés, à commencer par le recours à des ondes millimétriques, à des architectures Massive MIMO, ainsi qu'à la découpe du réseau (network slicing), ce dernier offrant le bénéfice de la meilleure allocation des ressources radio, ainsi que de l'adaptabilité de la qualité de service à l'usage de l'application. La 5G suscite, par ailleurs, des problèmes importants de gestion du spectre, de sécurité des communications, ainsi que de risque d'interférence, dans le contexte de la multiplication des antennes, ainsi que de la complexité de l'architecture. Je soutiens ainsi que la 5G doit être regardée non seulement comme une nouvelle norme radio, mais comme un levier central de la transformation numérique des systèmes productifs, des services publics, ainsi que, plus largement, de l'utilisation du numérique.

1.4 Les applications de la 5G

La technologie de télécommunication sans fil de cinquième génération est désignée sous le terme de nouvelle technologie radio sans fil (NR). Cette technologie devra satisfaire des cas d'utilisation répartis de la manière suivante :

• L'eMBB : ou Enhanced Mobile Broadband, désigne des communications mobiles à très haut débit. Cette utilisation s'inscrit dans la continuité des générations antérieures de téléphonie mobile et répond à l'accroissement exponentiel de l'usage des données mobiles. Les applications de cette catégorie sont généralement des flots vidéo de qualité supérieure ainsi que des applications de réalité virtuelle et augmentée. Par conséquent, l'objectif est de satisfaire une demande croissante en matière de volume de données et de rapidité de transfert [3].

• Le mMTC : pour la communication immense entre machines : les échanges entre objets. L'internet des objets appartient à cette catégorie. Cette catégorie concerne la capacité à gérer un nombre considérable de connexions, atteignant jusqu'à un million par kilomètre carré. La quantité de données à transmettre par communication est généralement restreinte, et la vitesse de transfert est peu contraignante. Un domaine d'application typique est la ville intelligente, intégrant des réseaux de capteurs pour la gestion de divers services [3].

• L'uRLLC : ou communication ultra fiable à faible latence, désigne des communications critiques où la fiabilité et le temps de réponse sont essentiels. Le véhicule autonome est l'application principale de cette catégorie, mais les communications des services de sécurité et d'urgence sont également impliquées. Il ne doit y avoir ni échec ni interruption de communication, et la transmission doit être effectuée dans les plus brefs délais [3].

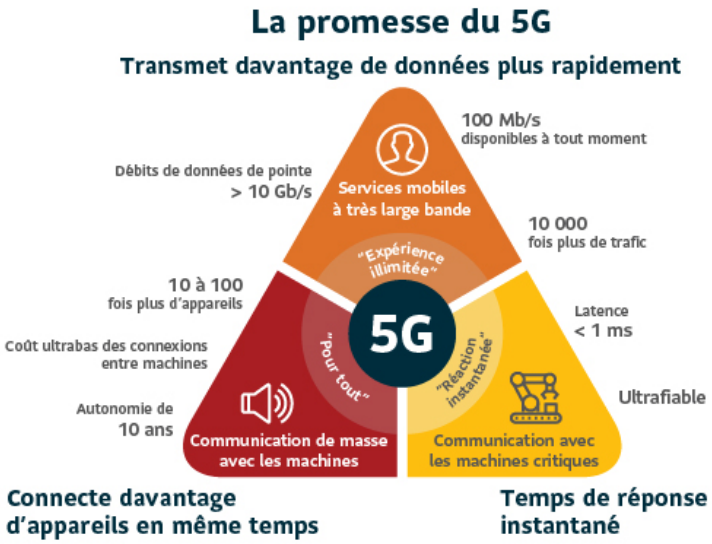


Fig. 1.2 – les cas d'utilisation du la 5G [4]

1.5 Bande fréquence (sub-6ghz - mmwave)

Les plages de fréquences utilisées par la technologie 5G. En ce qui concerne les bandes de fréquences exploitées, la 5G mobilise deux plages distinctes : la plage FR1, également nommée Sub-6 GHz, s'étendant de 410 MHz à 6 GHz, et la plage FR2, dite mmWave, couvrant la gamme de 6 GHz à 100 GHz, ces deux plages jouant des fonctions complémentaires. Pour la plage FR1, l'emploi de fréquences plus basses autorise une portée étendue ainsi qu'une meilleure pénétration à l'intérieur des bâtiments. Cependant, les bandes basses de la gamme FR1, exploitées par les opérateurs télécoms pour les réseaux 2G, 3G et 4G, demeurent également les plus sollicitées. La plage FR1 présente ainsi un risque de saturation en raison de débits insuffisamment élevés. La bande FR2 présente une capacité élevée et, par conséquent, une latence diminuée, cependant sa couverture demeure restreinte en raison de l'atténuation importante des ondes dans les milieux urbains. Les bandes de fréquences ciblées dans cette thèse correspondent aux bandes n° 78 (3,4-3,8 GHz) de la plage FR1 ainsi qu'à n° 258 (24,5-27,5 GHz) de la plage FR2 [5].

Actuellement, les fréquences Sub-6 GHz sont principalement exploitées dans les systèmes sans fil en raison d'une insuffisance notable de bandes de fréquences disponibles pour satisfaire les exigences des services sans fil. [6] La gamme de fréquences élevée millimétrique (mm-Wave) est considérée comme essentielle pour préserver l'efficacité spectrale et exploiter pleinement le potentiel des communications multi-antennes à large diffusion. Cependant, la communication en ondes millimétriques demeure intrinsèquement sujette à une faible fiabilité, caractérisée par des risques de congestion, d'importantes atténuations du signal et de balayage des canaux. Théoriquement, cette approche pourrait accroître la capacité du service de réseau à large bande mobile améliorée (eMBB) et diminuer le délai de transmission afin d'as-

surer un fonctionnement plus discret. En conséquence, il conviendrait d'intégrer de manière harmonieuse la souplesse des réseaux micro-ondes avec les performances des réseaux millimétriques afin d'assurer des communications ultra-fiables à faible latence (URLLC) ainsi qu'un accès sans fil à haut débit. Les communications par ondes millimétriques (mm-Wave) attirent une attention importante dans le domaine de la recherche en raison de l'importante largeur de bande disponible, permettant potentiellement des débits de plusieurs gigabits par seconde par utilisateur [6].

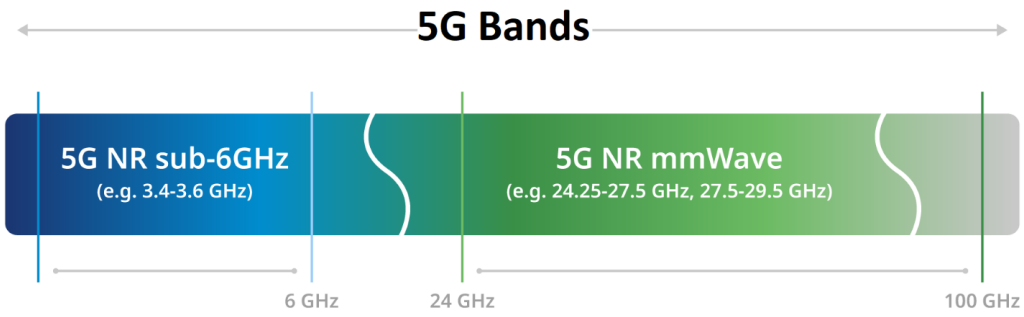


Fig. 1.3 – Répartition de la bande fréquence (sub-6ghz - mmwave) [4], [7]

1.6 Technologies clés de la 5G

1. New spectrum : La 5G constitue la première technologie de radiocommunication mobile développée pour opérer sur l'ensemble des bandes de fréquence comprises entre 400 MHz et 90 GHz. Les bandes de fréquences basses sont indispensables pour assurer la couverture, tandis que les bandes de fréquences hautes sont destinées à fournir des débits de données élevés ainsi qu'une capacité accrue. Les premières mises en œuvre de la 5G emploient la duplexation par répartition dans le temps (TDD) dans la plage de 2,5 à 5,0 GHz, la duplexation par répartition en fréquence (FDD) en dessous de 2,5 GHz, ainsi que la TDD aux fréquences millimétriques comprises entre 24 et 39 GHz [8].

2. Massive Multiple Input Multiple Output (MIMO) : Le beamforming est susceptible d'augmenter de façon importante l'efficacité spectrale et la couverture du réseau. Et se révèle plus efficace à des fréquences élevées. Ainsi, la dimension de l'antenne, proportionnelle à la longueur d'onde, diminue avec l'augmentation de la fréquence. En pratique, le MIMO massif peut être utilisé à des fréquences supérieures à 1 GHz au niveau des stations de base, tandis que les ondes millimétriques peuvent être directement utilisées dans les équipements. Le Massive MIMO sera intégré aux spécifications et aux déploiements initiaux de la 5G [8].

3. Network slicing : Les couches physique et protocolaire de la 5G doivent être flexibles afin d'accueillir des scénarios d'exploitation et des bandes de fréquence variés dans le but d'opti-

miser l'efficacité spectrale et celle de la consommation électrique. Le découpage du réseau (Network slicing) engendrera des sous-réseaux virtuels différents pour des services différents au sein d'un même réseau 5G. Ce découpage permet aux opérateurs de gérer plusieurs cas d'usage avec des clients entreprises sans avoir besoin de réseaux dédiés [8].

4. Dual connectivity and LTE co-existence : La 5G peut fonctionner comme système autonome (SA), mais dans la pratique les premiers déploiements s'appuieront largement sur une cohabitation avec la LTE existante, afin de profiter des infrastructures déjà en place. Les terminaux compatibles 5G peuvent établir en parallèle des liaisons radio avec les réseaux 5G et LTE. Cette double connectivité sert à lisser la transition en améliorant la continuité de service, en augmentant le débit disponible pour l'utilisateur final et en réduisant le risque de rupture de connexion lors des changements de cellule ou de technologie. Dans ce cadre, la 5G ne remplace pas immédiatement la LTE. Elle s'intègre plutôt à celle-ci, puisqu'elle a été spécifiquement pensée pour coopérer avec la technologie LTE, ce qui autorise le partage dynamique du spectre radio et simplifie les opérations de refarming et de réorganisation des bandes de fréquences au fil de l'évolution du réseau [8].

5. Support for cloud implementation and edge computing : L'architecture contemporaine des réseaux LTE repose sur une totale distribution au niveau radio et une intégrale centralisation dans le réseau central. Pour réduire la latence, il faut rapprocher le contenu du réseau radio, ce qui entraîne une sortie locale et le recours à l'informatique en périphérie. La scalabilité implique d'intégrer les avantages du cloud dans les réseaux radio grâce à une architecture de cloud en périphérie. La radio et le cœur 5G sont conçus pour être nativement déployés dans le cloud, et incluent notamment de nouvelles interfaces au cœur du réseau radio [8].

6. Beamforming : Le beamforming constitue une approche prometteuse pour optimiser les performances des réseaux mobiles. Le façonnage de faisceau permet d'obtenir une efficacité spectrale accrue, ce qui accroît considérablement la capacité des sites de stations de base existants. Le traitement par formation de faisceaux peut, en outre, optimiser les performances de la liaison tout en assurant une couverture spatiale élargie. Le beamforming est un concept étudié dans la littérature académique depuis plusieurs années, cependant, il n'a pas été intégré dans les premières spécifications LTE de la Release 3GPP 8. Le beamforming a été intégré aux spécifications 5G dès la version initiale 3GPP Release 15. Les bénéfices du beamforming sont effectivement réalisés en pratique grâce aux antennes massives MIMO. L'objectif consiste à optimiser intégralement la conception radio 5G pour le beamforming dans les systèmes massifs MIMO. Le concept fondamental du beamforming est présenté dans la Fig. 1.4. La méthode conventionnelle diffuse les données sur l'ensemble de la zone cellulaire, alors que le beamforming les transmet à l'aide de faisceaux étroits ciblant les utilisateurs. Les ressources identiques peuvent être allouées à plusieurs utilisateurs dans un même secteur, ce qui permet de réduire les interférences et d'accroître la capacité de la cellule.

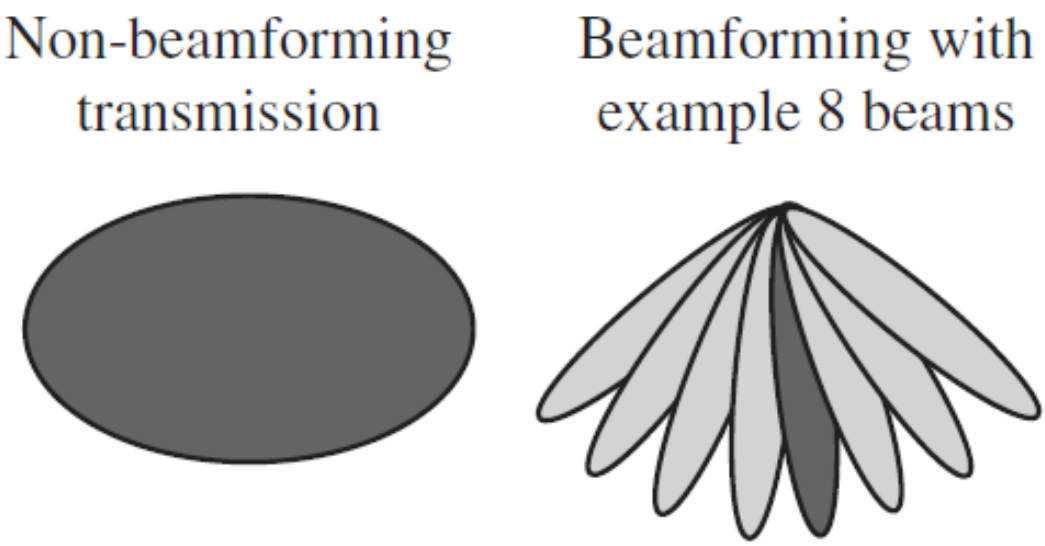


Fig. 1.4 – Éléments fondamentaux de la technologie 5G [8].

1.7 Architecture du réseau 5G

À la différence des générations antérieures de réseaux mobiles, le déploiement de la 5G n'exige pas la mise en place concomitante du réseau central 5G (5GC) et du réseau d'accès radio de nouvelle génération (NG-RAN). [9] Les architectures NSA (Accès Non-Standalone) et SA (Accès Standalone) constituent deux modèles distincts de réseaux 5G : – Le SA constitue une architecture de cœur de réseau intégralement nouvelle, fondée sur les services : chaque nœud radio est gouverné de manière autonome par le réseau central 5G. Une architecture orientée services propose des fonctionnalités sous la forme d'un ensemble de Fonctions Réseau (FN). Les fonctions réseau (NFs) au sein de l'architecture centrale de la 5G sont conçues pour être nativement intégrées au cloud. [9]– Le mode NSA s'appuie soit sur le réseau central 4G, soit sur le réseau central 5G. La NSA connecte le signal de contrôle au réseau central par l'intermédiaire d'un nœud maître radio (Master Node). Le MN correspond soit à un nœud radio 4G, soit à un nœud radio 5G. Le mobile node (MN) gère un secondary node (SN) (nœud radio 4G ou nœud radio 5G) conformément au mécanisme de double connectivité (DC). [9] L'architecture 5G autonome requiert la mise en place d'un réseau central 5G interfacé avec le NG-RAN. [9] Les architectures 5G-NSA et 5G-SA ont été définies dans la Release 15 de la spécification 3GPP [9].

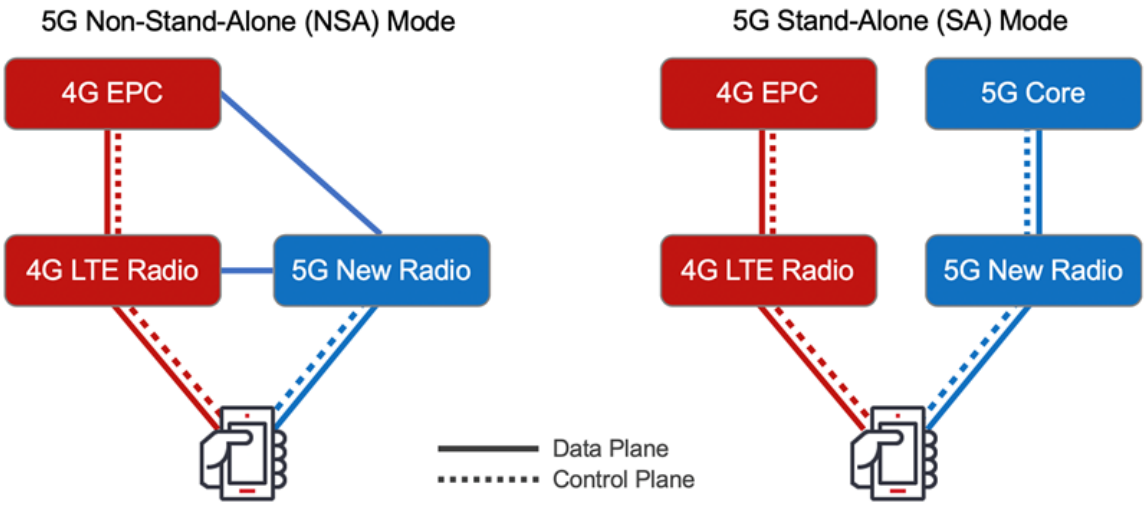


Fig. 1.5 – 5G NSA, 5G SA [10].

La figure 1.7 présente une représentation simplifiée de l'architecture du réseau 5G. Les stations de base correspondent soit à des gNB assurant les terminaisons des protocoles du plan utilisateur et du plan de contrôle, soit à des ng-eNB. Les gNB représentent les stations de base 5G natives intégrant par défaut les fonctionnalités radio 5G, tandis que les ng-eNB correspondent à des infrastructures 4G améliorées afin de prendre en charge les services radio 5G, dans le but de faciliter une transition progressive entre les deux normes.[11] Les stations gNB et ng-eNB sont reliées entre elles au moyen de l'interface Xn. L'ensemble de ces stations forme le réseau d'accès de nouvelle génération (NG-RAN, Next Generation-Radio Access Network). L'interconnexion des gNB et ng-eNB au réseau cœur 5G (5GC) s'établit via les fonctions réseau AMF (Access and Mobility Management Function) et UPF (User Plane Function) du cœur, par l'intermédiaire de l'interface NG [11].

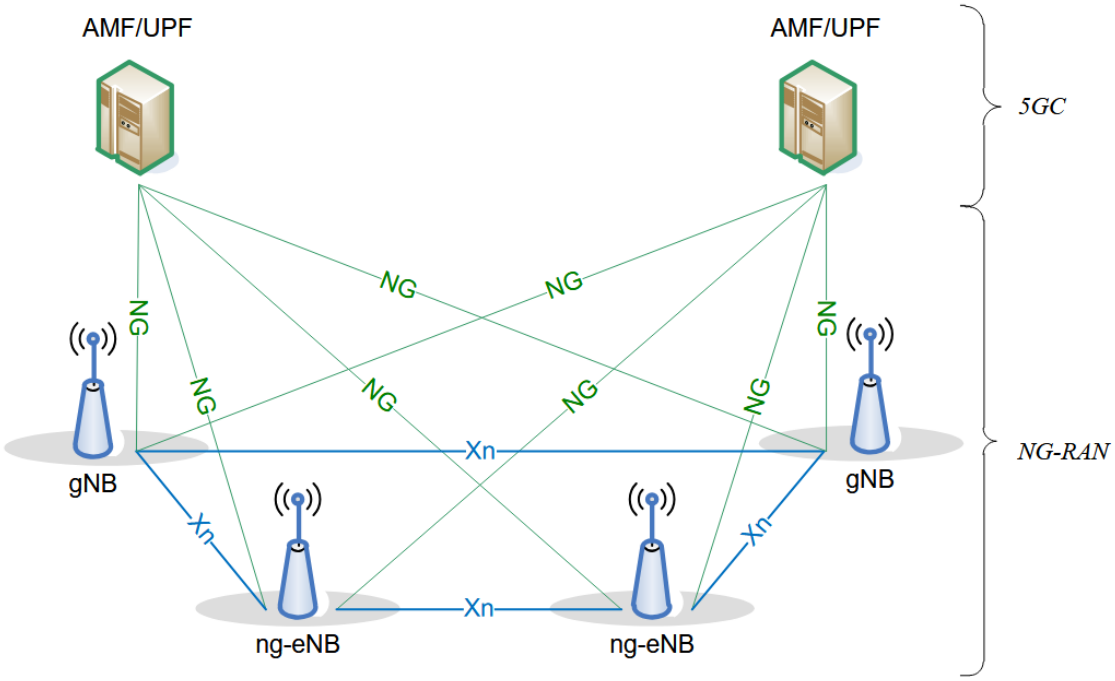


Fig. 1.6 – Architecture du réseau 5G [11]

La fonction AMF est responsable de la gestion de la signalisation indépendante des données utilisateur, notamment en matière de mobilité et de sécurité. Elle prend en charge des terminaux utilisateurs (UE, User Equipment) présentant divers profils de mobilité. L'UPF assure la gestion des fonctionnalités relatives aux données utilisateur, notamment le routage des paquets, la qualité de service (QoS) et le reporting du trafic utilisateur.[11] L'AMF et l'UPF constituent les fonctions du réseau central qui interagissent avec le réseau d'accès. Les fonctions complémentaires du réseau central sont illustrées à la figure 1.8, laquelle présente l'architecture fonctionnelle du réseau 5G. Une synthèse succincte de ces fonctions est présentée dans le tableau 1-1[11].

Tab. 1.1 – Architecture fonctionnelle du réseau 5G [12].

Acronyme	**Nom complet**	**Définition**
NRF	*Network Repository Function*	Fournit le support pour la gestion des services des fonctions réseau : enregistrement, autorisation et découverte.
NEF	*Network Exposure Function*	Fournit l'accès externe aux capacités des fonctions réseau, notamment la surveillance, la supervision et l'exposition sécurisée des services.
UDM	*Unified Data Management*	Assure le stockage des données et permet la séparation entre les fonctions de traitement et les fonctions de stockage.
NSSF	*Network Slice Selection Function*	Assure la sélection et la gestion du *Network Slicing*.
AUSF	*Authentication Server Function*	Assure la gestion de l'authentification des utilisateurs.
PCF	*Policy Control Function*	Fournit un cadre unifié pour le contrôle du comportement du réseau.
SMF	*Session Management Function*	Assure la gestion des sessions ainsi que l'allocation et la gestion des adresses IP des équipements utilisateurs, appelés UEs.
DN	*Data Network*	Gère l'interaction avec les réseaux de données externes.
AF	*Application Function*	Fonction spécifiée et déployée par l'opérateur pour fournir des services spécifiques.

Du point de vue architectural, le déploiement de la 5G s'effectue en deux phases, illustrées à la figure 1.9 [13]. La première étape correspond à un déploiement assisté ou non autonome (NSA, Non-Stand Alone). Un réseau d'accès hybride LTE-ngRAN est interconnecté à un réseau cœur LTE amélioré afin de prendre en charge simultanément les fonctions fondamentales 5G (5G EPC, cœur de paquet évolué 5G) et les services LTE (options 1 et 3). Cette architecture emploie la technologie LTE en tant qu'entité responsable de la gestion du plan de contrôle, tandis que la gestion du plan utilisateur est assurée conjointement par les deux normes. Cette double connectivité facilitera la prise en charge des débits 5G tout en assurant une couverture LTE améliorée. Cette architecture progressera vers un déploiement autonome (SA, Stand-Alone) dans lequel le réseau central sera un réseau natif 5G. Cependant, les fonctionnalités radio LTE demeureront supportées (options 5 et 7).

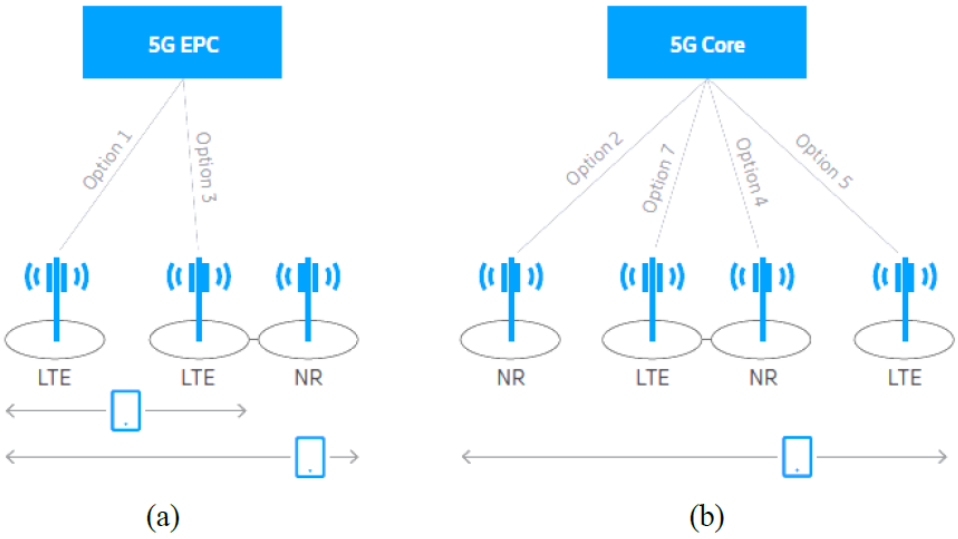


Fig. 1.7 – Modes de déploiement autonome (a) et non-autonome (b) du réseau 5G [13]

1.8 Les interférences dans les réseaux cellulaires

Dans le cadre de la communication cellulaire, l'interférence se caractérise par l'ajout d'un signal parasite à un signal utile. Fondamentalement, il s'agit d'une altération perturbatrice d'un signal. L'interférence se distingue du bruit, lequel peut correspondre à tout élément perturbant le signal utile. Le bruit peut résulter de la température, des signaux vocaux intenses émis par les machines, des rayons gamma, des impuretés, entre autres facteurs. Par conséquent, il ne s'agit pas d'une interférence ; toutefois, celle-ci altère les performances du réseau ainsi que l'expérience utilisateur. Dans le réseau cellulaire de nouvelle génération comportant plusieurs nœuds à faible puissance, un déploiement aléatoire ainsi que des scénarios de réutilisation de fréquence, le système engendre d'importantes problématiques d'interférence. Par conséquent, l'interférence constitue le principal enjeu dans la conception innovante d'un nouveau spectre exploitable pour les systèmes cellulaires 5G contemporains. [14].

1.8.1 Définition des interférences

L'interférence, c'est la présence de signaux radio indésirables qui perturbent la réception du signal attendu. Cette interférence peut supprimer le signal prévu, provoquer une perte de signal ou affecter la qualité du son ou de l'image. On trouve généralement comme principales causes d'interférence les émetteurs et les appareils électriques. Cette interférence peut être continue ou intermittente, provenant de la réfraction ou de la diffraction à la surface de la terre ou dans les couches atmosphériques (troposphère). Les interférences peuvent réduire la qualité du signal, augmenter le délai (latence) et diminuer la vitesse de transmission des

données (débit). Dans des cas plus sérieux, les interférences peuvent même conduire à des pannes de communication.

Les interférences sont des perturbations qui surviennent du fait du fonctionnement d'autres systèmes de communication.

Les interférences vont influencer la quantité de puissance du signal qu'un récepteur reçoit. La quantité d'interférence dépendra de la distance entre le système récepteur et le système émetteur ainsi que d'autres facteurs, tels que les conditions météorologiques et environnementales [15].

1.8.2 Types d'interférences

Interférence Inter-Cellulaire (IIC)

Une source classique de problèmes dans les réseaux mobiles est l'interférence entre cellules, souvent due à l'utilisation partagée des mêmes ressources par les cellules voisines. Cette interférence peut réduire les performances du réseau dans son ensemble et affecter la qualité du signal pour les utilisateurs situés en bordure.

Le déploiement étendu de stations de base à faible puissance et simplicité accrue au cœur de la couverture macro cellulaire, couplé à la réutilisation des maigres ressources spectrales, a conduit à un renforcement marqué de l'interférence inter-cellulaire (IIC). [16]

Interférence intra-couche et inter-couche

Bien que les réseaux à multiples niveaux accroissent significativement la capacité du réseau, ils se heurtent à des défis majeurs, en particulier en ce qui concerne la gestion des interférences entre les cellules adjacentes. De manière générale, deux catégories d'interférences inter-cellulaires sont susceptibles de dégrader les performances des réseaux hétérogènes : l'interférence intra-couche entre les entités du réseau appartenant à la même couche, ainsi que l'interférence inter-couche (également désignée sous le terme d'interférence entre couches) entre les entités du réseau issues de couches différentes, telles que l'interférence entre les cellules petites (Small Cell, SC) et les cellules macros (Macro Cell, MC). [16]

Interférence de Canal Adjacent (ICA)

De manière générale, l'interférence inter-cellulaire résulte soit de l'exploitation du même canal, désignée sous le terme d'interférence co-canal (CCI), soit du chevauchement entre canaux adjacents, qualifiée d'interférence de canal adjacent (ICA).

L'ICA résulte principalement de filtres récepteurs défectueux, qui laissent passer des fréquences adjacentes dans la bande passante, ainsi que de la non-linéarité des amplificateurs.

Un enjeu majeur des réseaux OFDM réside dans les émissions hors bande intenses, génératrices d'interférences significatives. Étant donné que la coexistence des canaux adjacents constitue un problème fréquent dans les réseaux mobiles, les autorités de régulation ont d'ores et déjà mis en place des normes rigoureuses relatives aux fuites des canaux adjacents ainsi qu'aux émissions parasites dans la bande opérationnelle afin de restreindre l'interférence de canal adjacent (ICA). Par exemple, les effets de l'interférence co-canal (ICC) peuvent être atténués par une sélection rigoureuse du filtre passe-bande au niveau du récepteur, l'emploi de schémas de modulation avancés ainsi que de techniques de traitement du signal, l'attribution d'une bande de protection et l'allocation de canaux adjacents à des cellules situées dans des régions géographiques distinctes [16].

Interférence Co-Canal (ICC)

L'interférence inter-cellulaire co-canal constitue la catégorie d'interférence la plus largement analysée dans la littérature scientifique. Dans un réseau FDD ou TDD statique et strictement synchronisé, les investigations scientifiques se focalisent sur l'interférence intra-mode. Par ailleurs, l'interférence entre modes (ou interférence croisée) survient généralement dans les réseaux TDD dynamiques ou en duplex intégral [16].

Interférence Intra-mode (ICI Descendant et Ascendant)

Dans le contexte de l'interférence inter-cellulaire co-canal en mode intra, désignée comme mode intra, deux catégories d'interférences sont identifiables selon la direction de la transmission :

- Interférence intercellulaire en liaison montante : l'équipement utilisateur constitue la source d'agression en générant une interférence sur la liaison montante d'une station de base adjacente, qui en subit les effets.
- Interférence intercellulaire en liaison descendante : la station de base constitue la source de l'interférence en liaison descendante affectant un équipement utilisateur (UE) provenant d'une cellule adjacente [16].

1.9 Paramètres radio utilisés pour l'analyse

SINR

Le rapport signal sur interférence plus bruit (SINR) se définit comme la proportion entre la puissance du signal et la somme de la puissance moyenne des interférences provenant des autres cellules ainsi que du bruit de fond [17].

RSSI

L'indicateur de puissance du signal reçu par le porteur (RSSI) correspond à la moyenne linéaire de la puissance totale reçue (exprimée en [W]) mesurée exclusivement au sein de certains symboles OFDM des sous-trames de mesure, dans la bande passante de mesure, sur un nombre N de blocs de ressources, par l'UE, en provenance de toutes les sources, incluant les cellules de service et hors-service en co-canal, l'interférence de canaux adjacents, le bruit thermique, etc [18].

RSRP

(Reference Signal Received Power) ou La puissance du signal de référence reçue constitue l'un des paramètres de performance essentiels au sein des réseaux LTE et 5G FWA. Cela renvoie exclusivement à l'intensité du signal reçu provenant de la cellule de service. Typiquement, des niveaux satisfaisants de RSRP se situent entre -85 dBm et -110 dBm pour la LTE, ainsi qu'entre -85 dBm et -115 dBm pour la 5G. Des niveaux de puissance du signal de référence reçus avoisinant 0 dBm correspondent à un signal plus intense, impliquant ainsi une meilleure couverture, alors que des valeurs inférieures (nombres négatifs plus élevés en valeur absolue) peuvent suggérer des difficultés liées à un signal faible ou une distance accrue entre l'utilisateur et la station de base. Ces paramètres peuvent être affectés par la dimension cellulaire, le milieu environnant ainsi que les caractéristiques techniques de l'appareil utilisé. C'est un paramètre indispensable pour décider du handover [19].

RSRQ

Le RSRQ, référence Signal Received Quality ou Signal de référence reçu Qualité, correspond à un indicateur de qualité qui intègre à la fois le niveau d'interférences et le bruit présents sur le canal lors de la réception du signal de référence. Il ne décrit donc pas uniquement la puissance reçue, mais reflète plutôt la manière dont les ressources radio sont réellement exploitées pour fournir le service. Cette mesure dépend fortement du contexte de déploiement du réseau, par exemple de la densité de stations de base, de l'architecture retenue, des choix de configuration radio ou encore des exigences propres au service ciblé, comme la voix sur LTE ou les usages nécessitant une très faible latence, où la stabilité du lien radio devient déterminante. Dans les réseaux LTE et 5G, on admet en général qu'un RSRQ compris entre -3 dB et -10 dB reste dans une plage jugée satisfaisante, avec des valeurs pouvant être encore meilleures lorsque l'environnement radio est particulièrement favorable. En règle générale, des valeurs plus élevées, donc moins négatives, traduisent une meilleure qualité de signal, et les opérateurs fixent un seuil minimal de RSRQ à respecter afin de garantir la stabilité des sessions, de maintenir un débit utile suffisant et de préserver la qualité d'expérience perçue par les utilisateurs. C'est un paramètre indispensable pour décider du handover [19].

CQI

En ce qui concerne l'exploitation des ressources, un paramètre essentiel est l'indicateur de qualité du canal (CQI).

Channel quality indicator. Où l'indicateur de qualité du canal constitue essentiellement une mesure de la qualité du canal radioélectrique utilisé pour la communication entre l'équipement utilisateur (UE) et la station de base (eNodeB en LTE, gNodeB en 5G). Il évalue la qualité du signal ainsi que le degré d'interférence perçu par l'UE. L'attribution des blocs de ressources (RB) dans les réseaux LTE et 5G nécessite généralement la prise en considération des conditions du canal, lesquelles sont généralement caractérisées par des métriques telles que le rapport signal sur interférence plus bruit (SINR) dans les réseaux LTE ou l'indicateur de qualité de canal (CQI) dans les réseaux 5G.

Une plage usuelle de valeurs CQI jugées acceptables pour un réseau 5G s'étend approximativement de 1 à 15 ou davantage, une valeur plus élevée indiquant une meilleure qualité du canal. Des indices de qualité de canal (CQI) élevés indiquent que les conditions du canal sont favorables ; par conséquent, les débits de données ainsi que l'efficacité spectrale peuvent être améliorés.

BLER

Le BLER se définit comme le rapport entre la quantité de blocs erronés et le nombre total de blocs réceptionnés sur le canal sans fil [20].

1.10 Effets des interférences sur les performances du réseau (QoS)

Dans un réseau cellulaire, notamment dans les systèmes récents comme la 5G NR, les interférences sont l'une des limites majeures aux performances réelles du service. Elles affectent directement la qualité de service (QoS) perçue par l'usager, en dégradant à la fois le débit, la latence et la fiabilité globale des transmissions.

1.10.1 Dégradation du débit

Les interférences présentes abaissent le rapport signal sur interférence plus bruit (SINR). Si le SINR se dégrade, l'émetteur doit utiliser des schémas de modulation et de codage plus conservateurs, comme passer d'une modulation 64-QAM à une QPSK, pour garder un taux d'erreur acceptable. Cette adaptation se traduit par une réduction du débit instantané, une perte d'efficacité spectrale et donc une diminution de la capacité globale de la cellule. En

pratique, l'utilisateur voit alors ses téléchargements ralentir, ses flux vidéo se couper, et son expérience réseau se dégrader de façon visible.

1.10.2 Augmentation de la latence

Les interférences provoquent également plus d’erreurs de transmission, ce qui active des mécanismes de retransmission comme le HARQ. Chaque diffusion introduit un retard, auquel il faut ajouter les temps de traitement pour détecter et corriger les erreurs. On constate donc une augmentation du temps de réponse du réseau et de la latence de bout en bout. Effet direct sur l'utilisateur : des performances médiocres pour les services sensibles au temps, comme les jeux en ligne, les conférences vidéo ou certaines applications IoT critiques. Parfois même inutilisables ;

1.10.3 Augmentation du taux d’erreur

En dégradant la qualité du signal reçu, les interférences élèvent le BLER (Block Error Rate) et plus généralement le taux d’erreurs de transmission. Cette situation entraîne des pertes de paquets, la nécessité de répéter de nombreuses transmissions et, finalement, une forte diminution de la fiabilité du réseau pour les services qui requièrent une continuité et une stabilité élevées.

Chapitre 2

Machine Learning appliqué aux réseaux 5G



Résumé :
Ce chapitre présente le Machine Learning comme une brique essentielle de la supervision intelligente des réseaux 5G. Il combine une approche théorique, une orientation télécom et une lecture orientée data. Le chapitre introduit d'abord les notions fondamentales de l'apprentissage automatique, puis explique comment les données issues des réseaux mobiles - KPI radio, journaux de stations de base, traces du coeur 5G, mesures de trafic et indicateurs de qualité de service - peuvent être transformées en variables exploitables par des modèles de classification, de prédiction ou de détection d'anomalies. Les différentes familles de méthodes sont ensuite détaillées : apprentissage supervisé, apprentissage non supervisé, apprentissage profond et apprentissage par renforcement. Les algorithmes les plus utiles pour la détection d'interférences et l'optimisation réseau, notamment la régression logistique, Random Forest, XGBoost, SVM, KNN, CNN et LSTM, sont expliqués avec leurs principes, leurs équations principales, leurs avantages et leurs limites. Une attention particulière est accordée au pipeline data télécom, au choix des métriques, à l'interprétabilité et à l'intégration opérationnelle dans des architectures 5G avancées, O-RAN et zero-touch. Le chapitre prépare ainsi la partie expérimentale du mémoire, en montrant comment passer d'une mesure radio brute à une décision réseau exploitable.

2.1 Introduction au Machine Learning dans les réseaux 5G

Le Machine Learning, ou apprentissage automatique, est une branche de l'intelligence artificielle qui permet à un système d'apprendre à partir des données afin de produire des prédictions, des classifications ou des décisions sans être explicitement programmé pour chaque situation. Contrairement aux méthodes classiques fondées uniquement sur des règles fixes, il s'appuie sur l'identification de régularités statistiques dans les données. Un modèle entraîné peut ensuite généraliser ces régularités à de nouvelles observations [21].

Dans le domaine des télécommunications, cette capacité est devenue particulièrement importante. Les réseaux mobiles modernes génèrent une grande quantité de données : indicateurs radio mesurés par les équipements utilisateurs, compteurs de performance des stations de base, traces de mobilité, journaux du coeur réseau, mesures de qualité de service et flux de trafic. Ces données contiennent des informations précieuses sur l'état réel du réseau. Les exploiter manuellement devient difficile, car les conditions radio changent rapidement selon la mobilité, la charge, la propagation, les obstacles, les interférences et le type de service demandé [22, 23].

Avec la 5G, le besoin d'automatisation augmente encore. Les réseaux 5G doivent prendre en charge des services hétérogènes comme l'eMBB, le mMTC et l'uRLLC. Ils doivent aussi gérer le Massive MIMO, le beamforming, le slicing, l'edge computing et des architectures de plus en plus virtualisées. Dans ce contexte, le Machine Learning devient un outil d'aide

à la décision pour plusieurs fonctions : détection d'interférences, prédiction de congestion, optimisation de la puissance, sélection de cellule, handover, allocation de ressources radio, maintenance prédictive et supervision automatisée.

L'intérêt du Machine Learning dans ce mémoire est donc double. D'une part, il permet d'analyser des mesures 5G afin de distinguer un état normal d'un état perturbé par des interférences. D'autre part, il permet de replacer cette détection dans une logique plus large de supervision intelligente du réseau. Les travaux de normalisation récents confirment cette orientation : 3GPP étudie l'AI/ML pour le NR et le NG-RAN, notamment pour la collecte de données, l'optimisation de mobilité, l'équilibrage de charge et la gestion d'énergie [24]. De même, l'architecture O-RAN introduit les RIC, xApps et rApps pour permettre des optimisations intelligentes du RAN [25], tandis que l'ETSI ZSM vise une gestion automatisée de bout en bout des réseaux et services [26].

2.2 Données télécoms et problématique data

Dans une application télécom, le Machine Learning ne commence pas par l'algorithme, mais par la donnée. Le réseau 5G produit plusieurs familles de données qui peuvent être utilisées pour construire un modèle :

- les **KPI radio**, tels que SINR, RSRP, RSRQ, RSSI, CQI, BLER, débit, latence, puissance reçue et taux de retransmission ;
- les **données de trafic**, telles que volume de données, nombre d'utilisateurs actifs, charge cellule, occupation des ressources physiques et files d'attente ;
- les **traces de mobilité**, telles que handover, vitesse estimée, cellule servante, cellule voisine et historique de changement de cellule ;
- les **logs réseau**, issus du RAN, du coeur 5G ou des systèmes d'exploitation et de maintenance ;
- les **données de supervision**, utilisées pour suivre la QoS, la QoE, les alarmes, les erreurs et les événements inhabituels.

Ces données sont souvent hétérogènes : certaines sont continues, d'autres catégorielles ; certaines sont échantillonnées régulièrement, d'autres sont événementielles ; certaines sont propres à l'utilisateur, d'autres à la cellule ou au coeur réseau. Le rôle du pipeline data consiste donc à transformer cette matière brute en un jeu de données cohérent, propre et exploitable.

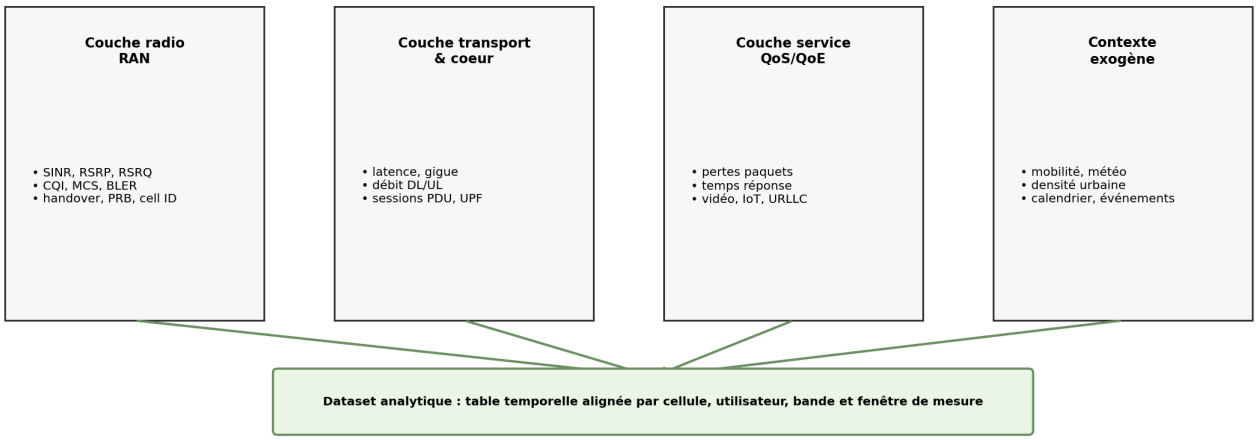


Fig. 2.1 – Sources de données télécoms exploitables pour l'apprentissage automatique dans un réseau 5G.

Dans le cadre de la détection d'interférences, une mesure réseau peut être représentée par un vecteur de caractéristiques :

$$\mathbf{x}_i = [\mathrm{SINR}_i,\ \mathrm{RSRP}_i,\ \mathrm{RSRQ}_i,\ \mathrm{RSSI}_i,\ \mathrm{CQI}_i,\ \mathrm{BLER}_i,\ R_i,\ \tau_i,\ L_i]^T \in \mathbb{R}^d, \tag{2.1}$$

où R_i représente le débit, τ_i la latence et L_i un indicateur de charge réseau. Pour une classification binaire, l'étiquette associée peut être définie par :

$$y_i = \begin{cases} 0, & \text{fonctionnement normal du réseau,} \\ 1, & \text{présence d'interférence ou de perturbation.} \end{cases} \tag{2.2}$$

Le jeu de données complet s'écrit alors :

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}. \tag{2.3}$$

L'objectif général est d'apprendre une fonction $f_{\boldsymbol{\theta}}$ capable d'associer à chaque vecteur de KPI une décision :

$$\hat{y}_i = f_{\boldsymbol{\theta}}(\mathbf{x}_i). \tag{2.4}$$

2.3 Pipeline data et apprentissage automatique

Le pipeline complet peut être vu comme une chaîne allant de la mesure réseau vers la décision opérationnelle. Cette chaîne comprend la collecte, l'ingestion, le nettoyage, la normalisation, l'extraction de caractéristiques, la séparation des données, l'entraînement, l'évaluation, l'in-

terprétation et éventuellement le déploiement.

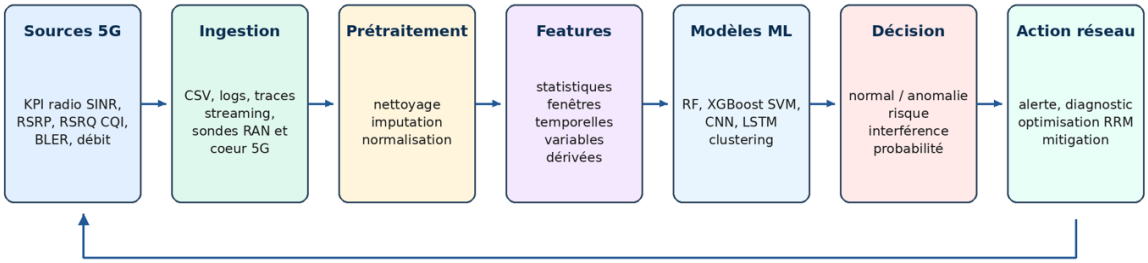


Fig. 2.2 – Pipeline data complet pour passer des mesures 5G à une décision réseau assistée par Machine Learning.

2.3.1 Collecte et ingestion des données

La collecte constitue la phase initiale. Dans un réseau 5G réel, les données peuvent provenir de sondes radio, de smartphones de mesure, de journaux de stations gNB, de traces de coeur réseau, d'outils de supervision ou de plateformes de simulation. Pour la détection d'interférences, il est important que les données couvrent à la fois des conditions normales et des conditions dégradées : baisse de SINR, dégradation de RSRQ, augmentation du BLER, chute du débit, hausse de latence ou instabilité de cellule.

L'ingestion consiste à centraliser ces données dans une structure exploitable. Dans une architecture data moderne, les données peuvent être stockées sous forme de fichiers CSV, bases SQL, bases NoSQL, data lake ou flux streaming. Pour un mémoire expérimental, un fichier tabulaire est souvent suffisant. Pour une application opérateur, il faut plutôt considérer des flux continus et des mécanismes de monitoring.

2.3.2 Prétraitement et nettoyage

Le prétraitement vise à rendre les données compatibles avec les algorithmes. Il comprend la suppression des doublons, le traitement des valeurs manquantes, la correction d'incohérences, la conversion des unités, l'encodage de variables catégorielles et la mise à l'échelle des variables numériques.

La normalisation centrée-réduite d'une caractéristique x_{ij} est donnée par :

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{2.5}$$

avec :

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \qquad \sigma_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2}. \tag{2.6}$$

Cette étape est particulièrement importante pour SVM, KNN et les réseaux de neurones, car ces méthodes sont sensibles à l'échelle des variables. Une normalisation min–max peut aussi

être utilisée :

$$x_{ij}^{(\mathrm{norm})} = \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_j^{\min}}. \tag{2.7}$$

2.3.3 Feature engineering télécom

Le feature engineering consiste à créer des variables plus informatives que les mesures brutes. Dans un réseau 5G, on peut créer des indicateurs comme la moyenne glissante du SINR, la variance du RSRQ, le rapport débit/charge cellule, la variation temporelle du CQI, ou encore un indicateur combiné de dégradation radio.

Pour une séquence de SINR mesurée sur une fenêtre de longueur M, la moyenne, la variance et la valeur quadratique moyenne sont :

$$\overline{\mathrm{SINR}} = \frac{1}{M} \sum_{n=0}^{M-1} \mathrm{SINR}[n], \tag{2.8}$$

$$\sigma^2_{\mathrm{SINR}} = \frac{1}{M-1} \sum_{n=0}^{M-1} \left(\mathrm{SINR}[n] - \overline{\mathrm{SINR}}\right)^2, \tag{2.9}$$

$$\mathrm{RMS}_{\mathrm{SINR}} = \sqrt{\frac{1}{M} \sum_{n=0}^{M-1} \mathrm{SINR}^2[n]}. \tag{2.10}$$

Une baisse de la moyenne du SINR combinée à une hausse de sa variance peut indiquer une situation radio instable.

Lorsque des mesures fréquentielles sont disponibles, le signal discret $x[n]$ peut être analysé par transformée de Fourier discrète :

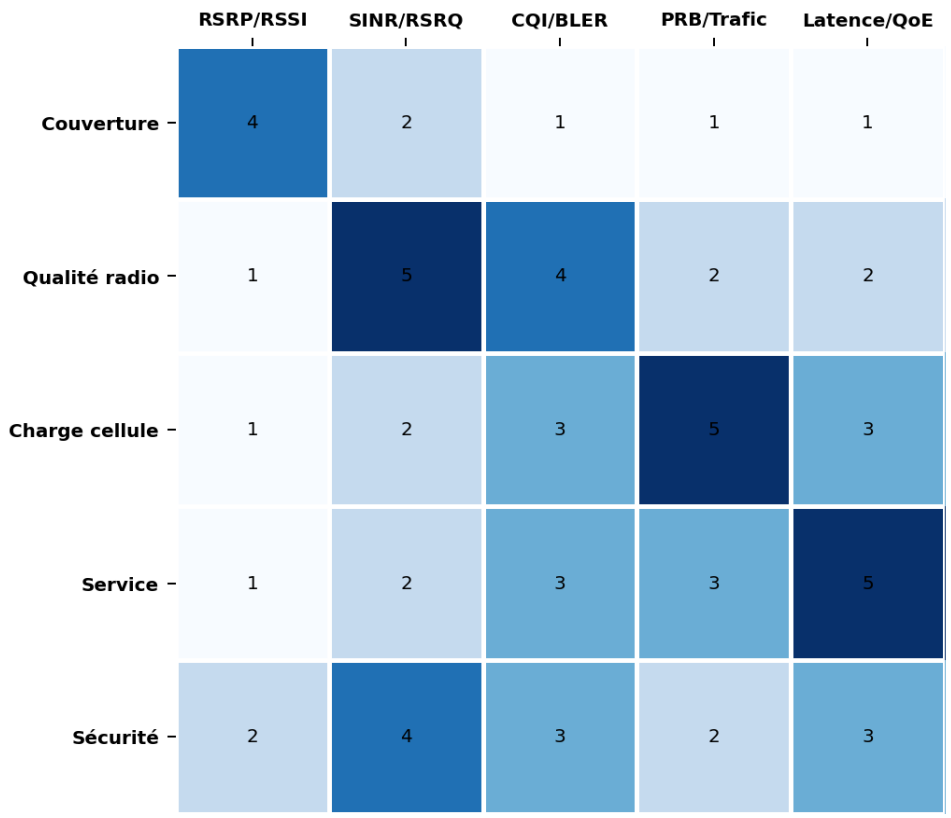
$$X[k] = \sum_{n=0}^{M-1} x[n] e^{-j2\pi kn/M}, \qquad k = 0, 1, \ldots, M-1. \tag{2.11}$$

La puissance spectrale estimée est alors :

$$P[k] = \frac{|X[k]|^2}{M}. \tag{2.12}$$

Un indicateur d'énergie d'interférence dans une bande fréquentielle $\mathcal{B}$ peut être défini par :

$$\eta_{\mathcal{B}} = \frac{\sum_{k \in \mathcal{B}} P[k]}{\sum_{k=0}^{M-1} P[k]}. \tag{2.13}$$



Exemple de matrice de pertinence : 1 = faible, 5 = très fort pour la détection d'interférences et l'analyse QoS.

Fig. 2.3 – Exemples de transformation des KPI 5G en caractéristiques utilisables par un modèle d'apprentissage.

2.3.4 Entraînement, validation et test

L'entraînement correspond à la phase pendant laquelle l'algorithme apprend une relation entre les variables d'entrée et la sortie attendue. En classification, le modèle cherche à séparer des observations normales et perturbées. En régression, il peut prédire un débit, une latence ou une probabilité de congestion. En détection d'anomalies, il peut apprendre le profil normal du réseau et signaler les écarts.

La séparation des données doit éviter les fuites d'information entre apprentissage et test. Une stratégie classique consiste à utiliser un ensemble d'entraînement, un ensemble de validation pour régler les hyperparamètres et un ensemble de test final. Dans le cas de données temporelles, il faut être prudent : un découpage aléatoire peut mélanger des observations trop proches dans le temps. Un découpage chronologique ou par session de mesure est souvent plus réaliste.

L'estimation des paramètres d'un modèle peut être formulée comme une minimisation du

risque empirique :

$$\boldsymbol{\theta}^{\star} = \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y_i, f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right) + \lambda \Omega(\boldsymbol{\theta}) \right], \tag{2.14}$$

où $\mathcal{L}$ est la fonction de perte, $\Omega(\boldsymbol{\theta})$ un terme de régularisation et λ un coefficient qui contrôle le compromis entre erreur et complexité.

2.3.5 Évaluation des performances

Après l'entraînement, les performances sont évaluées à partir de la matrice de confusion. Pour une classification binaire, on définit les vrais positifs (TP), faux positifs (FP), vrais négatifs (TN) et faux négatifs (FN). Les métriques usuelles sont :

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2.15}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2.16}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2.17}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.18}$$

Dans une application de détection d'interférences, le rappel est critique, car un faux négatif correspond à une interférence présente mais non détectée. Cependant, un excès de faux positifs peut aussi être problématique, car il peut déclencher des actions inutiles de maintenance ou d'optimisation. Le choix de la métrique principale dépend donc du contexte opérationnel.

2.4 Typologie des méthodes de Machine Learning

Les méthodes de Machine Learning peuvent être regroupées en plusieurs familles. Chacune répond à un besoin différent dans les télécommunications : classification d'états radio, détection d'anomalies, prédiction de trafic, optimisation de ressources ou décision séquentielle.

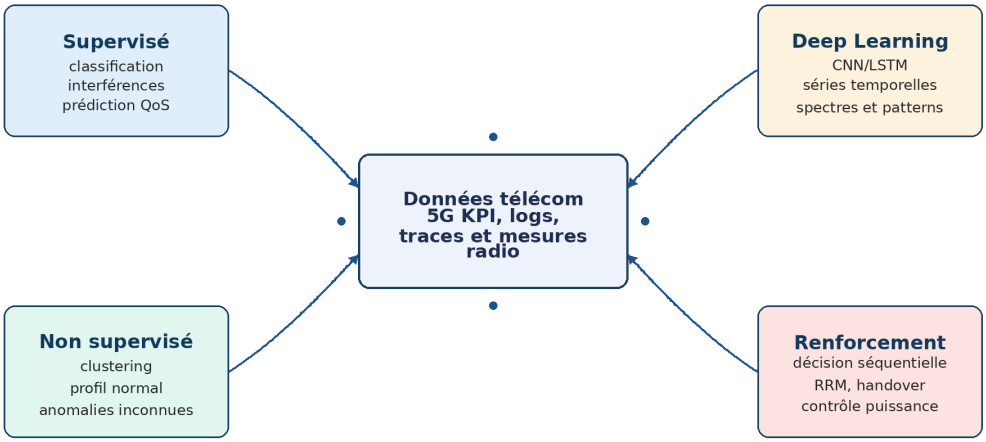


Fig. 2.4 – Correspondance entre familles de méthodes Machine Learning, types de données télécoms et usages 5G.

2.4.1 Apprentissage supervisé

L'apprentissage supervisé utilise des données étiquetées. Chaque observation $\mathbf{x}_i$ est associée à une sortie connue y_i. Dans ce mémoire, cette sortie peut représenter l'état du réseau : normal ou perturbé. L'objectif est d'apprendre une fonction :

$$f : \mathbb{R}^d \to \mathcal{Y}, \tag{2.19}$$

où $\mathbb{R}^d$ est l'espace des caractéristiques et $\mathcal{Y}$ l'ensemble des classes.

Pour une classification à C classes, la sortie peut être représentée sous forme probabiliste :

$$\mathbf{p}_i = [P(y_i = 1|\mathbf{x}_i), P(y_i = 2|\mathbf{x}_i), \ldots, P(y_i = C|\mathbf{x}_i)]\,. \tag{2.20}$$

La classe prédite est alors :

$$\hat{y}_i = \arg \max_{c \in \{1,\ldots,C\}} P(y_i = c|\mathbf{x}_i). \tag{2.21}$$

Dans les réseaux 5G, l'apprentissage supervisé convient aux situations où des labels sont disponibles : interférence co-canal, brouillage, fonctionnement normal, congestion ou défaut de configuration. Il est particulièrement adapté à la classification d'états radio à partir des KPI.

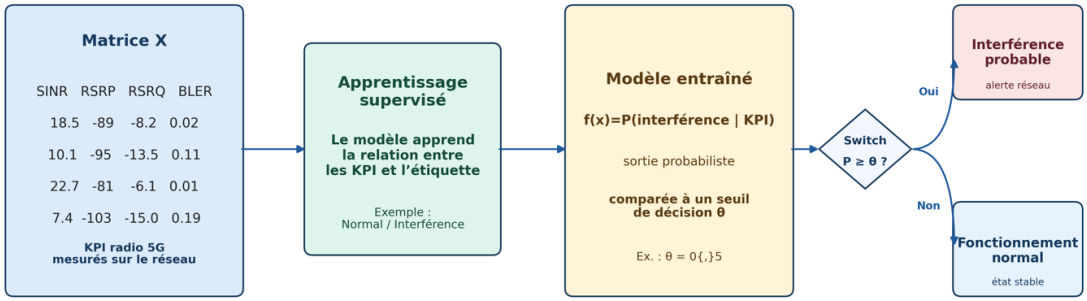


Fig. 2.5 – Principe d'une classification supervisée d'états réseau à partir de KPI 5G.

2.4.2 Apprentissage non supervisé

L'apprentissage non supervisé ne nécessite pas de labels. Il cherche à découvrir des structures cachées dans les données. Dans les réseaux 5G, il est utile lorsque les anomalies sont rares, mal connues ou non annotées. Un modèle peut apprendre le comportement normal du réseau, puis signaler les observations qui s'en éloignent.

Pour un algorithme de clustering comme K-means, l'objectif est de partitionner les observations en K groupes en minimisant :

$$J = \sum_{i=1}^{N} \min_{k \in \{1,\dots,K\}} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 , \tag{2.22}$$

où $\boldsymbol{\mu}_k$ est le centre du cluster k. Une observation peut être considérée comme anormale si sa distance au centre le plus proche dépasse un seuil γ :

$$s_i = \min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2 , \qquad s_i > \gamma \Rightarrow \text{anomalie}. \tag{2.23}$$

Cette approche est pertinente pour détecter des comportements inattendus : brouillage inconnu, changement de propagation, cellule mal configurée ou dérive progressive de la qualité radio.

2.4.3 Apprentissage profond

L'apprentissage profond repose sur des réseaux de neurones multi-couches capables de modéliser des relations non linéaires complexes. Pour une couche ℓ, on écrit :

$$\mathbf{h}^{(\ell)} = \phi\left(\mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)}\right), \tag{2.24}$$

où $\mathbf{W}^{(\ell)}$ est la matrice des poids, $\mathbf{b}^{(\ell)}$ le biais et ϕ une fonction d'activation. Pour une classification, la fonction softmax est souvent utilisée :

$$P(y = c|\mathbf{x}) = \frac{\exp(a_c)}{\sum_{r=1}^{C} \exp(a_r)}. \tag{2.25}$$

La perte d'entropie croisée est :

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log \left(P(y_i = c|\mathbf{x}_i) \right). \tag{2.26}$$

Pour des séries temporelles de KPI, une convolution 1D peut extraire des motifs locaux :

$$h_m[t] = \phi \left(\sum_{c=1}^{C_{\text{in}}} \sum_{r=0}^{K-1} w_{m,c,r} x_c[t - r] + b_m \right). \tag{2.27}$$

Cette formulation est utile pour détecter des transitions rapides dans le SINR, le RSRQ ou le débit. Les architectures LSTM et GRU sont aussi adaptées à la prédiction de trafic, à la détection de dérive et à l'analyse de séquences de mesures.

2.4.4 Apprentissage par renforcement

L'apprentissage par renforcement est utilisé lorsqu'un agent doit prendre une suite de décisions dans un environnement dynamique. Dans un réseau 5G, l'état s_t peut représenter la charge cellule, le SINR moyen, le nombre d'utilisateurs, le taux de handover ou le BLER. L'action a_t peut correspondre à un réglage de puissance, une réallocation de ressources ou une décision de handover. La récompense r_t mesure la qualité de la décision, par exemple un compromis entre débit, latence, stabilité et énergie.

Le but est d'apprendre une politique $\pi(a|s)$ qui maximise le retour cumulé :

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{2.28}$$

où γ est un facteur d'actualisation. Cette famille est prometteuse pour l'optimisation RRM, mais elle demande une modélisation prudente, car une mauvaise action peut dégrader le réseau réel. Elle est donc souvent étudiée d'abord en simulation ou dans des environnements contrôlés.

2.5 Algorithmes de classification et d'analyse utilisés en télécom

2.5.1 Régression logistique

La régression logistique est un modèle simple et interprétable pour la classification binaire. Elle calcule la probabilité d'interférence à partir de la fonction sigmoïde :

$$P(y_i = 1|\mathbf{x}_i) = \sigma\left(\mathbf{w}^T\mathbf{x}_i + b\right), \tag{2.29}$$

avec :

$$\sigma(u) = \frac{1}{1+e^{-u}}. \tag{2.30}$$

La décision est :

$$\hat{y}_i = \begin{cases} 1, & P(y_i = 1|\mathbf{x}_i) \geq \alpha, \\ 0, & P(y_i = 1|\mathbf{x}_i) < \alpha. \end{cases} \tag{2.31}$$

La fonction de coût est la perte logistique :

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]. \tag{2.32}$$

Dans un contexte 5G, la régression logistique est utile comme modèle de référence. Elle permet de vérifier rapidement si les KPI contiennent une information discriminante. Ses coefficients peuvent aussi donner une première indication sur le poids relatif du SINR, du RSRQ ou du BLER dans la décision.

2.5.2 Random Forest

Random Forest combine plusieurs arbres de décision entraînés sur des sous-échantillons des données. Chaque arbre produit une décision, puis la prédiction finale est obtenue par vote majoritaire :

$$\hat{y} = \arg\max_{c}\sum_{t=1}^{T}\mathbb{I}\left(h_t(\mathbf{x}) = c\right). \tag{2.33}$$

À chaque noeud, une séparation est choisie pour réduire l'impureté. L'indice de Gini est :

$$G = 1 - \sum_{c=1}^{C} p_c^2. \tag{2.34}$$

Le gain d'une séparation est :

$$\Delta G = G_{\text{parent}} - \left(\frac{N_L}{N} G_L + \frac{N_R}{N} G_R \right). \tag{2.35}$$

Ce modèle est très pertinent pour les données tabulaires télécoms. Il tolère le bruit, gère les relations non linéaires et fournit une importance des variables. Dans la détection d'interférences, il peut montrer si les décisions reposent principalement sur le SINR, le RSRQ, le BLER ou des variables dérivées.

2.5.3 XGBoost

XGBoost est un algorithme de gradient boosting qui construit progressivement une série d'arbres faibles. Chaque nouvel arbre corrige les erreurs résiduelles des arbres précédents. Sa fonction objectif peut s'écrire :

$$\mathcal{J}^{(t)} = \sum_{i=1}^{N} \mathcal{L} \left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i) \right) + \Omega(f_t), \tag{2.36}$$

où f_t est le nouvel arbre ajouté et $\Omega(f_t)$ un terme de régularisation. XGBoost est souvent très performant sur les données tabulaires, car il capture des interactions complexes tout en contrôlant le surapprentissage.

Dans les réseaux 5G, il peut être utilisé pour la classification d'anomalies, la prédiction de QoS, la priorisation d'alertes et l'estimation du risque d'interférence. Sa rapidité d'inférence le rend intéressant pour des systèmes de supervision proches du temps réel.

2.5.4 SVM

Le SVM cherche un hyperplan qui sépare les classes en maximisant la marge. Dans le cas linéaire :

$$\mathbf{w}^T \mathbf{x} + b = 0. \tag{2.37}$$

Avec marge souple, le problème devient :

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{N} \xi_i, \tag{2.38}$$

sous les contraintes :

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0. \tag{2.39}$$

Avec un noyau RBF :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \right). \tag{2.40}$$

Le SVM est utile lorsque la frontière entre classes est complexe, mais il peut devenir coûteux sur de très grands jeux de données. Il demande aussi une normalisation rigoureuse des variables.

2.5.5 K plus proches voisins

KNN classe une observation selon les classes de ses k voisins les plus proches. La distance euclidienne est :

$$d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{d} (x_j - x_{ij})^2}. \tag{2.41}$$

La classe prédite est :

$$\hat{y} = \arg\max_{c} \sum_{i \in \mathcal{N}_k(\mathbf{x})} \mathbb{I}(y_i = c). \tag{2.42}$$

KNN est simple et pédagogique, mais sensible à l'échelle des variables et au volume des données. Dans les télécoms, il peut servir d'approche exploratoire, mais il est moins adapté aux systèmes temps réel de grande taille.

2.5.6 CNN et LSTM pour séries temporelles de KPI

Lorsque les KPI sont observés dans le temps, les modèles séquentiels deviennent utiles. Un CNN 1D peut détecter des motifs locaux comme une chute rapide du SINR ou une augmentation brève du BLER. Un LSTM peut mémoriser des dépendances plus longues, par exemple une dégradation progressive avant une congestion.

Dans un LSTM, l'état mémoire permet de conserver une information sur le passé. Cette propriété est utile pour les réseaux mobiles, car une interférence réelle se manifeste souvent comme une dynamique temporelle plutôt qu'un simple point isolé. Toutefois, ces modèles nécessitent davantage de données, un entraînement plus coûteux et une validation rigoureuse pour éviter le surapprentissage.

Cas télécom	Données	Méthodes adaptées	Avantage principal
Interférences	KPI radio étiquetés SINR, RSRQ, BLER	Random Forest XGBoost, SVM	classification robuste importance des variables
Anomalies inconnues	logs non étiquetés trafic, compteurs	K-means Isolation Forest Autoencodeur	découverte sans vérité terrain
Prédiction trafic	séries temporelles débit, charge cellule	LSTM, GRU Random Forest régression	anticipation congestion
Optimisation RRM	état réseau + action récompense QoS	Apprentissage par renforcement	décision dynamique
Beam management	mesures spatiales CSI, faisceaux	CNN, DNN modèles légers	réduction de mesures et overhead

Fig. 2.6 – Choix pratique des méthodes Machine Learning selon les cas d'usage télécoms et la nature des données disponibles.

2.6 Application à la détection des interférences 5G

La détection des interférences est une application centrale du Machine Learning en 5G. Une interférence peut provoquer une baisse de SINR, une dégradation du RSRQ, une augmentation du BLER, une chute de débit, une instabilité de handover et une augmentation de latence. Les modèles de Machine Learning peuvent apprendre à reconnaître ces signatures à partir de combinaisons de KPI.

Le SINR s'écrit :

$$\mathrm{SINR} = \frac{P_s}{P_I + P_N}, \tag{2.43}$$

où P_s est la puissance du signal utile, P_I la puissance d'interférence et P_N la puissance de bruit. Lorsque P_I augmente, le SINR diminue. Le RSRQ peut être écrit :

$$\mathrm{RSRQ} = \frac{N_{\mathrm{RB}} \cdot \mathrm{RSRP}}{\mathrm{RSSI}}. \tag{2.44}$$

Une dégradation simultanée du SINR et du RSRQ peut indiquer une forte occupation radio ou une interférence.

Un indicateur simple de risque d'interférence peut être défini comme :

$$\mathcal{I} = \alpha_1 \left(\frac{\mathrm{SINR}_{\mathrm{ref}} - \mathrm{SINR}}{\mathrm{SINR}_{\mathrm{ref}}} \right) + \alpha_2 \left(\frac{\mathrm{BLER}}{\mathrm{BLER}_{\max}} \right) + \alpha_3 \left(\frac{R_{\mathrm{ref}} - R}{R_{\mathrm{ref}}} \right), \tag{2.45}$$

avec $\alpha_1 + \alpha_2 + \alpha_3 = 1$. La décision peut être formulée par :

$$\mathcal{I} > \gamma \Rightarrow \text{présence probable d'interférence.} \tag{2.46}$$

Dans une approche Machine Learning, cet indicateur peut être utilisé comme variable supplémentaire, comme règle de comparaison ou comme base d'étiquetage lorsque les labels terrain ne sont pas disponibles.

2.7 Machine Learning pour l'optimisation des réseaux 5G

Au-delà de la détection, le Machine Learning peut contribuer à l'optimisation du réseau. Le SINR d'un utilisateur u connecté à une station de base b peut être approximé par :

$$\mathrm{SINR}_{u,b} = \frac{P_b G_{u,b}}{\sum_{b' \neq b} P_{b'} G_{u,b'} + N_0 B}, \tag{2.47}$$

où P_b est la puissance émise, $G_{u,b}$ le gain de canal, N_0 la densité spectrale de bruit et B la bande passante. Le débit peut être approximé par :

$$R_u = B_u \log_2 \left(1 + \mathrm{SINR}_{u,b}\right). \tag{2.48}$$

L'allocation des ressources peut être formulée comme :

$$\max_{\{B_u, P_u\}} \sum_{u=1}^{U} U_u(R_u), \tag{2.49}$$

sous contraintes :

$$\sum_{u=1}^{U} B_u \leq B_{\mathrm{tot}}, \qquad 0 \leq P_u \leq P_{\max}. \tag{2.50}$$

Dans la pratique, résoudre ce problème exactement en temps réel peut être difficile. Les modèles de Machine Learning peuvent apprendre à approximer de bonnes décisions à partir de l'état courant du réseau. Cela permet d'améliorer l'équilibrage de charge, la sélection de cellule, la mobilité, la gestion d'énergie ou l'allocation de ressources.

2.8 Architecture data, MLOps et boucle fermée

Pour être exploitable dans un réseau réel, un modèle Machine Learning doit être intégré dans une architecture data et MLOps. Cette architecture couvre le stockage des données, l'entraînement, la validation, le suivi des performances, le déploiement et la surveillance de dérive.

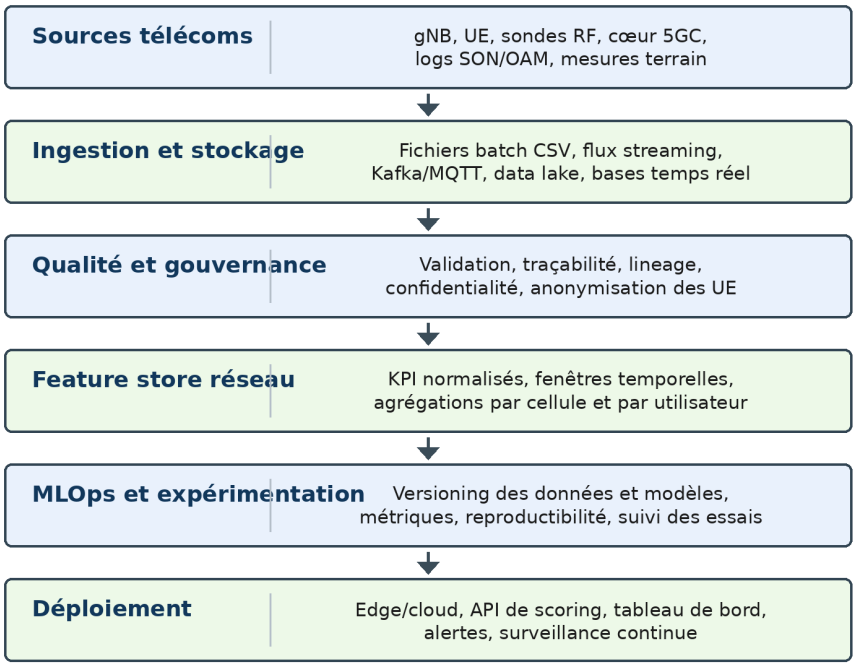


Fig. 2.7 – Architecture data et MLOps pour l'exploitation des modèles Machine Learning dans un contexte télécom.

Dans une approche opérationnelle, le modèle peut être intégré dans une boucle fermée : observer, analyser, décider, agir, vérifier. Cette logique rejoint les architectures d'automatisation des réseaux modernes. L'ETSI ZSM vise une gestion automatisée de bout en bout des réseaux et services [26]. L'O-RAN introduit également des composants comme le Non-RT RIC, le Near-RT RIC, les rApps et les xApps pour supporter des optimisations intelligentes du RAN [25].

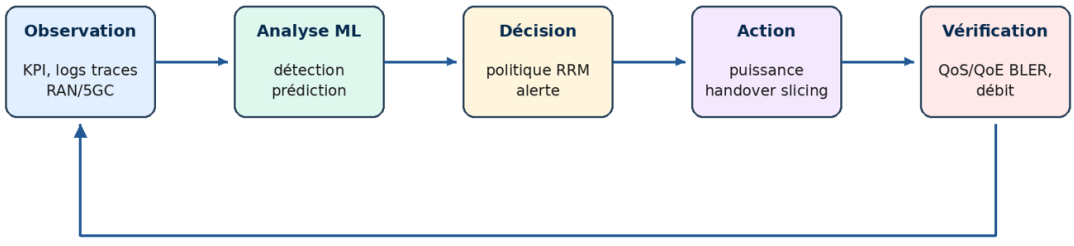


Fig. 2.8 – Principe d'une boucle fermée pour la supervision et l'optimisation intelligente d'un réseau 5G.

Dans ce cadre, la détection d'interférences ne constitue pas une fin en soi. Elle doit idéalement déclencher une action : alerte, changement de configuration, réallocation de ressources, modification de puissance, coordination inter-cellules ou diagnostic approfondi. Le modèle doit donc être évalué non seulement par ses métriques de classification, mais aussi par son impact sur la qualité de service.

2.9 Comparaison des méthodes ML pour les réseaux 5G

Tab. 2.1 – Comparaison des principales méthodes Machine Learning dans les applications télécoms 5G.

Méthode	Données adaptées	Utilisation en 5G	Avantages et limites
Régression logistique	KPI tabulaires normalisés	Modèle de référence pour distinguer normal/interférence	Simple, rapide, interprétable ; limité pour relations très non linéaires.
Random Forest	Données tabulaires bruitées	Détection d'anomalies, importance des KPI, classification d'états réseau	Robuste, performant, explicable ; modèle plus lourd qu'une régression simple.
XGBoost	Données tabulaires structurées	Prédiction QoS, risque de congestion, détection d'interférences	Très performant ; nécessite réglage soigneux des hyperparamètres.
SVM	Données de taille modérée	Classification avec frontière complexe	Efficace après normalisation ; coût élevé pour grands jeux de données.
KNN	Données exploratoires	Similarité entre états radio observés	Très simple ; sensible à la dimension, aux distances et au volume de données.
K-means / clustering	Données non étiquetées	Segmentation de profils réseau, détection d'anomalies inconnues	Ne nécessite pas de labels ; interprétation parfois délicate.
CNN 1D	Séries de KPI ou spectres	Détection de motifs locaux, signatures temporelles ou fréquentielles	Extraction automatique de motifs ; demande plus de données.
LSTM / GRU	Séries temporelles	Prédiction de trafic, dérive radio, évolution de QoS	Tient compte du passé ; coût d'entraînement plus élevé.
Renforcement	État-action-récompense	RRM, handover, contrôle de puissance, slicing	Décision dynamique ; validation nécessaire avant déploiement réel.

2.10 Revue synthétique des travaux existants

Les travaux récents confirment que l'intégration de l'AI/ML dans les réseaux 5G et 5G-Advanced constitue une direction majeure. Dans 3GPP, Release 18 a marqué une étape importante avec des études sur l'AI/ML pour l'interface NR, notamment le retour d'état de canal, la gestion de faisceaux et le positionnement [27, 28]. Les travaux 3GPP sur le NG-RAN portent également sur la collecte de données, l'équilibrage de charge, la mobilité et l'économie d'énergie [24].

Dans le domaine O-RAN, l'intelligence est intégrée au travers du Non-RT RIC et du Near-RT RIC. Les rApps peuvent produire des politiques d'optimisation à plus long terme, tandis que les xApps peuvent exécuter des ajustements plus proches du temps réel [25]. Cette architecture est particulièrement cohérente avec les applications de Machine Learning étudiées dans ce chapitre : prédiction, détection, optimisation et décision.

Enfin, l'ETSI ZSM propose une vision de gestion automatisée de bout en bout pour les réseaux et services [26]. Cette orientation est importante pour replacer la détection d'interférences dans une logique de système complet : un modèle détecte une situation anormale, mais la valeur opérationnelle apparaît lorsqu'une action corrective est déclenchée, vérifiée et améliorée dans le temps.

Tab. 2.2 – Synthèse de travaux et initiatives liés à l'AI/ML dans les réseaux 5G.

Source	Orientation	Lien avec ce chapitre	Apport principal
3GPP AI/ML pour NR	Interface radio intelligente	Beam management, CSI, positionnement, mobilité	Intégration progressive de l'AI/ML dans les standards 5G-Advanced.
3GPP AI/ML pour NG-RAN	Optimisation réseau	Collecte de données, load balancing, mobility optimization	Passage de la détection vers l'optimisation RAN.
O-RAN	RIC, xApps, rApps	Déploiement de modèles dans une architecture RAN ouverte	Support opérationnel des applications ML dans le RAN.
ETSI ZSM	Automatisation zero-touch	Boucles fermées et orchestration de bout en bout	Vision système pour l'automatisation réseau.
Travaux sur le jamming 5G	Détection d'interférences	Classification d'états radio et détection d'anomalies	Montre l'intérêt de modèles supervisés ou hybrides.

Source	Orientation	Lien avec ce chapitre	Apport principal
Travaux sur la propagation ML	Optimisation radio	Cartes radio, prédiction de couverture, planification	Exploitation des données pour améliorer la planification réseau.

2.11 Discussion et Limites des ML dans la 5G

L'utilisation du Machine Learning dans les réseaux 5G doit être comprise avec prudence. Un modèle performant sur un jeu de données donné n'est pas automatiquement robuste dans toutes les conditions terrain. Plusieurs points doivent être vérifiés : représentativité des données, qualité des labels, séparation correcte entraînement/test, dérive temporelle, diversité géographique, déséquilibre des classes et interprétabilité.

Dans la détection d'interférences, le problème des labels est particulièrement important. Si l'étiquette est construite à partir de seuils sur SINR et RSRQ, le modèle peut apprendre principalement cette règle plutôt qu'une vérité terrain indépendante. Cela reste utile comme preuve de concept, mais il faut le présenter comme une détection supervisée par règle et non comme une validation terrain complète. Une validation plus forte nécessiterait des mesures annotées par expertise, des scénarios contrôlés d'interférence ou des données opérateur.

Le choix du modèle doit aussi être adapté au besoin. Pour une première analyse tabulaire, Random Forest et XGBoost sont souvent de bons candidats. Pour une analyse temporelle, LSTM ou CNN 1D peuvent être pertinents. Pour des anomalies inconnues, clustering ou autoencodeurs peuvent être utilisés. Pour des décisions de contrôle, l'apprentissage par renforcement est prometteur, mais doit être testé avec prudence en simulation.

2.12 Conclusion

Ce chapitre a présenté une version complète et orientée télécom-data du Machine Learning appliqué aux réseaux 5G. Il a montré que l'apprentissage automatique ne se limite pas à un ensemble d'algorithmes, mais correspond à une chaîne complète allant de la donnée réseau brute vers la décision opérationnelle. Les KPI radio, les logs, les traces de mobilité et les compteurs de performance constituent la matière première. Le prétraitement, la normalisation, le feature engineering et l'évaluation assurent la qualité du modèle. Les méthodes supervisées, non supervisées, profondes et par renforcement répondent ensuite à différents besoins : classification, détection d'anomalies, prédiction, optimisation et décision dynamique.

L'accent a été mis sur l'application à la détection des interférences 5G, où les modèles peuvent

exploiter le SINR, le RSRP, le RSRQ, le BLER, le débit et la latence pour identifier des situations perturbées. Le chapitre a également relié cette problématique à des tendances actuelles comme l'O-RAN, les xApps/rApps, l'AI/ML pour le NG-RAN et les architectures zero-touch. Cette synthèse prépare la suite du mémoire, qui portera sur la construction expérimentale du jeu de données, le choix des modèles, leur implémentation et l'évaluation des performances obtenues.

Chapitre 3

Dataset et prétraitement

Présentation du chapitre

Après avoir présenté dans les chapitres précédents les fondements théoriques des réseaux 5G NR et les concepts généraux du Machine Learning, ce troisième chapitre constitue le cœur applicatif de notre travail. Il décrit l'ensemble de la chaîne de traitement permettant de détecter automatiquement les interférences à partir des indicateurs clés de performance (KPIs) normalisés par le 3GPP [29].

Nous y présentons successivement la problématique étudiée, le dataset utilisé [30], la procédure d'étiquetage fondée sur une modélisation physique simplifiée, ainsi que les différentes étapes de prétraitement des données indispensables à la phase d'apprentissage supervisé [31].

L'approche adoptée se veut à la fois rigoureuse sur le plan méthodologique et réaliste sur le plan opérationnel : les données exploitées proviennent de campagnes de mesures réelles en environnement 5G NR [30], tandis que la labellisation des classes d'interférence repose sur des seuils physiques documentés dans les spécifications 3GPP TS 38.215 [29] et dans les travaux récents de détection d'anomalies en réseau cellulaire [32, 33].

3.1 Description du problème étudié

Les réseaux 5G New Radio (5G NR) opèrent dans un environnement radio de plus en plus dense, partagé entre de nombreux opérateurs, technologies coexistantes (4G LTE, Wi-Fi 6, IoT) et bandes de fréquences sub-6 GHz et mmWave [29].

Dans ce contexte, les interférences constituent l'une des principales causes de dégradation de la qualité de service (QoS), entraînant une baisse du débit, une augmentation de la latence et des coupures de connexion [32].

La détection précoce des interférences est donc un enjeu majeur pour les opérateurs [32, 34]. Les méthodes traditionnelles, fondées sur des seuils statiques appliqués à un ou deux indicateurs (typiquement le SINR ou le RSRQ), souffrent de plusieurs limitations identifiées

dans la littérature [32, 33] :

Manque de finesse : ces méthodes ne capturent pas les interactions non linéaires entre les multiples KPIs.

Non-adaptative : les seuils fixes ne s'ajustent pas aux variations temporelles et spatiales du réseau.

Absence d'explicabilité : elles produisent une alerte binaire sans hiérarchiser les causes probables.

Difficulté à différencier les types d'interférence (co-canal, canal adjacent, externe).

Face à ces limitations, notre travail vise à concevoir un système intelligent de détection des interférences en réseau 5G NR exploitant les techniques de Machine Learning [35, 36] et de Deep Learning [37, 38].

La problématique centrale peut être formulée comme suit :

Comment, à partir d'un ensemble de KPIs radio (RSRP, RSRQ, RSNR, RSSI, CQI, débits descendant/montant), construire un modèle automatique capable de distinguer un état réseau normal d'un état perturbé par des interférences, tout en fournissant une explication interprétable de sa décision ?

Notre approche se décline en deux niveaux complémentaires :

Détection binaire : classification de chaque mesure en deux classes — Normal (0) ou Interférence (1).

Classification multi-classes : discrimination du type d'interférence (Aucune, Co-canal, Canal adjacent, Externe).

Cette démarche s'inscrit dans la philosophie des réseaux auto-organisés (SON — Self-Organizing Networks) et des architectures Open RAN intégrant nativement l'intelligence artificielle pour le monitoring et l'optimisation radio [32].

3.2 Présentation du dataset utilisé

3.2.1 Origine et nature du dataset

Pour mener à bien cette étude, nous avons exploité un jeu de données publié par Raca et al. En 2020 lors de la 11e ACM Multimedia Systems Conference [30]. Ce dataset, issu d'une campagne de mesures réelles effectuée sur un réseau commercial 5G NR d'un opérateur irlandais majeur, est, à notre connaissance, le premier ensemble public contenant à la fois des KPIs canaux et contextuels pour des connexions 5G [30]. Le fichier brut, intitulé B_2020_02_27_18_39_27.csv, est composé de 5 991 échantillons horodatés à la seconde, capturés par un équipement utilisateur (UE) compatible 5G durant une session de drive-test.

Les indicateurs radio (RSRP, RSRQ, SINR, CQI, RSSI, débits, etc.) Sont définis et mesurés

conformément à la spécification 3GPP TS 38.215 [29].

En revanche, les étiquettes de classe (Normal / Interférence et types d'interférence) ne figurent pas dans le fichier original. Elles ont été générées de manière synthétique à partir d'une modélisation physique fondée sur des seuils documentés dans les spécifications 3GPP [29] et la littérature scientifique récente en télécommunications [32, 33].

Cette approche hybride présente un double avantage :

Réalisme : les KPI reflètent les phénomènes radio observés sur un réseau opérationnel.

Maîtrise de l'étiquetage : les classes sont construites selon des règles physiquement justifiables et reproductibles.

3.2.2 Description desvariables

Le dataset comporte 26 colonnes regroupant des informations de géolocalisation, d'identification cellulaire et de qualité radio [30].

Pour les besoins de la détection d'interférences, nous avons retenu un sous-ensemble de variables directement pertinentes sur le plan physique, conformément aux définitions du 3GPP [29].

Le tableau3.1 présente les principales variables utilisées.

Tab. 3.1 – Variables principales utilisées pour la détection d'interférences

Variable	Unité	Signification physique
RSRP	dBm	*Reference Signal Received Power* : puissance moyenne reçue sur les signaux de référence. Il s'agit d'un indicateur clé de la couverture cellulaire[4] .
RSRQ	dB	*Reference Signal Received Quality* : qualité du signal de référence. Cette grandeur est sensible aux interférences, à la charge cellulaire et aux variations de la qualité radio [4].
SINR	dB	*Signal-to-Interference-plus-Noise Ratio : rapport signal sur interférence plus bruit. Ce* KPI est particulièrement discriminant pour détecter une dégradation due au bruit ou aux interférences [4].
RSSI	dBm	*Received Signal Strength Indicator* : puissance totale reçue sur la bande mesurée. Elle inclut le signal utile, les interférences et le bruit [4].

Variable	Unité	Signification physique
CQI	[1–15]	*Channel Quality Indicator* : indicateur entier rapporté à la station de base. Il traduit la qualité instantanée du canal et permet l'adaptation du schéma de modulation et de codage, appelé MCS [4].
DL_bitrate	Kbps	Débit descendant instantané mesuré au niveau de l'UE. Il reflète la performance globale de la liaison descendante et peut diminuer en présence d'interférences [3].
UL_bitrate	Kbps	Débit montant instantané. Il est utile pour détecter les dégradations affectant la liaison montante entre l'UE et le réseau [3].
NetworkMode	–	Type de réseau utilisé. Dans le jeu de données considéré, il correspond au mode 5G NR pour l'ensemble des échantillons [3].
CellID	–	Identifiant de la cellule serveuse. Cette information est utilisée pour contextualiser les mesures, mais elle n'est pas retenue comme variable d'entrée principale du modèle.
Timestamp	–	Horodatage des mesures, généralement disponible seconde par seconde. Il permet d'effectuer des analyses temporelles et de suivre l'évolution des KPI dans le temps.

Les colonnes liées au ping (PINGAVG, PINGMIN, PINGMAX, PINGSTDEV, PINGLOSS) contiennent une forte proportion de valeurs manquantes (plus de 92 %) et ont donc été écartées, conformément aux bonnes pratiques de gestion des données manquantes [39]. Les colonnes d'identification hexadécimale (CELLHEX, NODEHEX, LACHEX) ont également été exclues car non pertinentes pour la classification du phénomène physique d'interférence.

3.3 Génération du dataset

Comme indiqué précédemment, les classes d'interférence ne sont pas fournies nativement par les équipements de mesure [30].

Nous avons donc construit la variable cible à l'aide d'une modélisation physique simplifiée des conditions radio observées. Cette section décrit le raisonnement et les hypothèses retenues.

3.3.1 Modélisation physique simplifiée

La détection d'interférences en milieu radio repose sur l'observation conjointe de plusieurs grandeurs physiques définies par la norme 3GPP [29]. En présence d'un signal interférant, le rapport signal sur bruit (SINR)diminue significativement, la qualité du signal de référence (RSRQ) se dégrade, et l'indicateur de qualité du canal (CQI) rapporté par l'UE chute [32, 34]. À l'inverse, le niveau total reçu (RSSI) peut augmenter en raison de la superposition de signaux indésirables, tandis que la puissance utile (RSRP) reste relativement stable [29]. Cette signature physique permet de distinguer plusieurs scénarios.

Nous avons adopté une procédure d'étiquetage à deux niveaux.

étiquetage binaire (Normal / Interférence)

Une première règle, affinée après une analyse exploratoire des valeurs SHAP [40], consiste à considérer qu'une mesure correspond à un état d'interférence si :

SINR < 0 dB OU (SINR < 3 dB ET RSRQ < –16 dB)

Cette règle traduit deux situations physiques distinctes : un effondrement catastrophique du rapport signal sur bruit (cas évident d'interférence dominante), ou une dégradation combinée — moins extrême mais simultanée — du SINR et de la qualité du signal de référence, signature typique d'une interférence intra-système (co-canal ou canal adjacent) [32, 34].

b) Classification multi-classes (type d'interférence)

Pour aller plus loin, nous avons proposé une classification fine en quatre catégories, qui correspondent aux principales sources d'interférence rencontrées dans les déploiements 5G NR [32, 34, 33] :

Tab. 3.2 – Règles physiques utilisées pour l'étiquetage des classes d'interférences

Classe	**Règle de décision**
None (Normal)	SINR≥ 5 dB et $\mathrm{RSRQ} \geq -14$ dB
Co-Channel	$\mathrm{RSRP} > -95$ dBm et (SINR< 2 dB ou $\mathrm{RSRQ} < -17$ dB)
Adjacent Channel	$\mathrm{RSSI} > -85$ dBm et SINR< 3 dB
External	Tous les autres cas de dégradation non couverts par les règles précédentes

d'étiquetage multi-classes des interférences [29, 32, 34].

L'interprétation physique de chacune de ces classes est la suivante :

Co-canal : interférence provoquée par une cellule voisine émettant sur la même fréquence porteuse; le RSRP reste élevé (signal utile fort), mais le SINR ou le RSRQ chutent à cause de la superposition de signaux concurrents [34].

Canal adjacent : perturbation due à des émissions sur une fréquence proche (bande adjacente) qui débordent dans la bande utile. Le RSSI augmente fortement alors que le SINR reste bas [32].

Externe : interférence d'origine non cellulaire (équipements industriels, brouillage, fuites RF d'autres systèmes) qui dégrade le bilan radio sans correspondre aux signatures précédentes [33].

Aucune (None) : conditions radio saines avec marges de SINR et RSRQ confortables [29].

3.3.2 Hypothèses retenues

La construction du dataset repose sur un ensemble d'hypothèses simplificatrices qu'il convient d'expliciter pour garantir la transparence de la démarche :

H1 — Stationnarité locale : les conditions radio sont supposées stationnaires sur la durée d'une mesure (1 seconde), ce qui justifie l'utilisation directe des KPI instantanés sans lissage temporel préalable [30].

H2 — Indépendance des échantillons : chaque ligne du dataset est traitée comme un échantillon indépendant. Cette hypothèse, classique en classification supervisée [38], néglige les corrélations temporelles courtes mais reste valable pour les modèles tabulaires (Random Forest, XGBoost) [35, 36].

H3 - Universalité des seuils : les seuils utilisés pour l'étiquetage (SINR, RSRQ, RSSI, RSRP) sont considérés comme représentatifs des valeurs critiques généralement admises en 5G NR sub-6 GHz par la spécification 3GPP [29], bien qu'ils puissent être ajustés selon le contexte spécifique d'un opérateur.

H4 — Couverture cellulaire constante : le NetworkMode est exclusivement 5G dans l'intégralité du jeu [30], ce qui élimine la variabilité technologique et permet de se concentrer sur les phénomènes propres à la 5G NR.

H5 — Absence de mobilité extrême : la vitesse mesurée reste compatible avec un usage urbain ou semi-urbain (piéton ou véhicule à vitesse modérée) [30], ce qui exclut l'effet Doppler important rencontré en environnement à haute mobilité.

Ces hypothèses, bien que simplificatrices, sont compatibles avec un grand nombre de scénarios opérationnels et permettent d'établir une preuve de concept reproductible [32, 33]. Elles pourront être levées dans des extensions futures (intégration de modèles séquentiels, prise en compte de la mobilité, adaptation dynamique des seuils).

3.4 Prétraitement des données

La qualité de l'apprentissage automatique dépend de façon critique de la qualité des données fournies aux modèles [39, 38]. Nous décrivons dans cette section les quatre étapes successives de prétraitement appliquées : nettoyage des données, normalisation, séparation en ensembles d'entraînement et de test, puis sélection des variables (feature selection).

3.4.1 Nettoyage des données

L'inspection initiale du fichier brut a révélé plusieurs anomalies qu'il a fallu corriger avant toute exploitation :

Valeurs manquantes : certaines colonnes — notamment les indicateurs de ping — contiennent plus de 92 % de valeurs représentées par le caractère '–'. Étant donné le très faible taux d'observations exploitables, ces colonnes ont été entièrement écartées, conformément aux bonnes pratiques de data cleaning [39].

Encodage des absences : les valeurs manquantes représentées par '–' dans les colonnes NRxRSRP et NRxRSRQ ont été converties en NaN puis remplacées par la valeur médiane de la colonne correspondante, technique robuste aux valeurs aberrantes [39].

Suppression des doublons : quelques lignes redondantes (horodatage et coordonnées identiques) ont été supprimées afin d'éviter un biais d'apprentissage.

Vérification des plages physiques : chaque variable a été examinée pour s'assurer que ses valeurs restent dans des intervalles physiquement plausibles (par exemple, RSRP entre –140 et –40 dBm, CQI entre 0 et 15) conformément à 3GPP [29]. Aucune anomalie n'a été détectée après nettoyage.

Au terme de cette étape, le dataset exploitable comporte 5 991 échantillons et 7 KPIs radio principaux (RSRP, RSRQ, SINR, CQI, RSSI, DL_bitrate, UL_bitrate) [30, 29], auxquels viennent s'ajouter deux variables dérivées présentées en section 3.4.4.

Pour enrichir la représentation, deux variables dérivées (feature engineering) ont été calculées à partir des KPIs primaires [38] :

DL_UL_Ratio = DL_bitrate / (UL_bitrate + ε) — rapport entre les débits descendant et montant ; un déséquilibre extrême peut signaler une dégradation asymétrique de la liaison [32].

SNR_CQI_Product = SINR× CQI — produit traduisant une mesure composite de qualité radio, plus discriminante que chaque indicateur pris isolément.

3.4.2 Normalisation des données

Les variables retenues présentent des plages de valeurs très différentes : les débits peuvent atteindre plusieurs centaines de milliers de kbps, alors que le SINRvarie sur un intervalle

d'une trentaine d'unités seulement. Sans normalisation, les algorithmes sensibles à l'échelle (réseaux de neurones, SVM, KNN) seraient biaisés par les variables à grande amplitude, et la descente de gradient deviendrait moins efficace [37, 38].

Nous avons donc appliqué une standardisation Z-score via la classe StandardScaler de la bibliothèque Scikit-learn [41], qui transforme chaque variable selon la formule :

$x' = (x - \mu) / \sigma$

Où μ désigne la moyenne empirique de la variable et σ son écart-type. Après cette transformation, chaque feature est centrée (moyenne nulle) et réduite (écart-type unitaire) [39]. Cette normalisation est calculée uniquement sur l'ensemble d'entraînement, puis appliquée identiquement à l'ensemble de test, afin d'éviter toute fuite d'information (data leakage) [38].

3.4.3 Séparation des données d'entraînement et de test

L'évaluation honnête d'un modèle de Machine Learning exige une stricte séparation entre les données utilisées pour l'apprentissage et celles utilisées pour l'évaluation [38]. Nous avons divisé le dataset selon une proportion classique de 80 % pour l'entraînement et 20 % pour le test, en utilisant la fonction train_test_split de Scikit-learn [41] avec un échantillonnage stratifié.

Le mode stratifié (paramètre stratify=y) est essentiel ici : il garantit que la proportion des classes (Normal vs Interférence) est préservée dans chacun des deux sous-ensembles [41, 39]. Sans cette précaution, l'ensemble de test pourrait, par hasard, contenir trop peu d'exemples d'interférence pour produire des métriques fiables [42].

Un random_state=42 fixe garantit la reproductibilité de l'expérience. Le tableau 3.3 résume la répartition obtenue.

Tab. 3.3 – Répartition train/test du dataset 5G

Ensemble	**Échantillons**	**Proportion**	**Stratifié**
Entraînement (train)	≈ 4 793	80 %	Oui
Test	≈ 1 198	20 %	Oui
Total	5 991	100 %	—

Par ailleurs, l'analyse de la distribution des classes a révélé un déséquilibre : la classe Normal est nettement plus représentée que la classe Interférence [43]. Pour éviter qu'un modèle naïf n'obtienne une précision artificielle élevée en prédisant systématiquement la classe majoritaire, nous avons appliqué la technique SMOTE (Synthetic Minority Over-sampling Technique) proposée par Chawla et al. [43] et implémentée dans la bibliothèque imbalanced-learn [44], exclusivement sur l'ensemble d'entraînement. Cette méthode génère des exemples

synthétiques de la classe minoritaire par interpolation entre voisins, sans dupliquer les observations existantes [43]. L'ensemble de test reste inchangé, ce qui préserve la validité de l'évaluation [42].

3.4.4 Sélection des variables (Feature Selection)

La sélection des variables a pour objectif de ne conserver que les KPI réellement informatifs pour la tâche de classification, ce qui présente plusieurs avantages : réduction du coût calculatoire, amélioration de la généralisation et limitation du risque de surapprentissage [35, 38].

Notre démarche de sélection combine trois approches complémentaires :

Sélection par expertise métier : le choix initial des 7 KPIs (RSRP, RSRQ, SINR, CQI, RSSI, DL_bitrate, UL_bitrate) est guidé par leur signification physique reconnue dans la norme 3GPP TS 38.215 [29] et leur usage dans la littérature de l'analyse des réseaux radio [30, 32].

Analyse de corrélation : une matrice de corrélation de Pearson a été calculée pour identifier les variables redondantes [39]. Aucune paire ne présente une corrélation supérieure à 0,90, ce qui justifie de conserver l'ensemble des KPIs initiaux.

Importance des features par les modèles : après une première phase d'apprentissage, l'importance des variables a été estimée via trois techniques distinctes — l'importance native du Random Forest [35], les valeurs SHAP du modèle XGBoost [36, 40], et la moyenne des cartes Grad-CAM du réseau 1D-CNN [45]. Ces trois méthodes convergent vers les mêmes variables dominantes : SINR, RSRQ et CQI, conformément à l'intuition physique [29, 32].

Le jeu final de variables retenues pour l'entraînement comprend donc les 7 KPIs primaires auxquels s'ajoutent les 2 variables dérivées définies précédemment, soit 9 features au total [38]. Ce vecteur d'entrée constitue la base de toutes les expérimentations décrites dans le chapitre suivant.

3.5 Conclusion

Dans ce troisième chapitre, nous avons posé les fondations expérimentales de notre travail de détection des interférences en réseau 5G NR par apprentissage automatique. La problématique étudiée a été clairement formulée : il s'agit de dépasser les limites des seuils statiques traditionnels en concevant un système capable d'exploiter conjointement plusieurs indicateurs radio [29, 32], et de fournir non seulement une décision fiable mais aussi une explication interprétable [45, 40].

Nous avons ensuite présenté en détail le dataset utilisé, issu d'une campagne de mesures réelles sur un réseau commercial 5G publiée par Raca et al. [30], complété par une procédure d'étiquetage synthétique reposant sur une modélisation physique documentée par la norme

3GPP [29]. Les règles d'étiquetage binaire et multi-classes ont été explicitées, accompagnées des hypothèses simplificatrices retenues pour garantir la transparence et la reproductibilité de l'approche.

Enfin, la chaîne de prétraitement a été déroulée pas à pas : nettoyage des données [39], normalisation Z-score via Scikit-learn [41], séparation stratifiée train/test avec gestion du déséquilibre par SMOTE [43, 44], et sélection des variables guidée à la fois par l'expertise du domaine et par des analyses statistiques. Au terme de ce travail préparatoire, nous disposons d'un dataset propre, équilibré et bien dimensionné, prêt à alimenter les algorithmes de Machine Learning et de Deep Learning.

Le chapitre suivant sera consacré à la conception, à l'entraînement et à l'évaluation comparative des trois modèles retenus — Random Forest [35], XGBoost [36] et 1D-CNN [37] — ainsi qu'à l'interprétation de leurs décisions à l'aide des techniques d'explicabilité SHAP [40] et Grad-CAM [45].

Chapitre 4

Implémentation, workflows et résultats expérimentaux

4.1 Introduction

Ce chapitre présente la réalisation expérimentale complète du système de détection des interférences dans un réseau 5G par apprentissage automatique. Il transforme la méthodologie générale décrite précédemment en une chaîne reproductible, organisée autour d'un notebook Python. L'objectif n'est pas uniquement de rapporter des valeurs de performance, mais de montrer comment les données sont importées, comment les classes sont construites, comment les variables sont préparées, comment les modèles sont entraînés, puis comment leurs décisions sont interprétées.

L'approche suivie repose sur trois familles de modèles : Random Forest, XGBoost et un réseau de convolution unidimensionnel, noté 1D-CNN. Les deux premiers modèles appartiennent aux méthodes d'arbres ensemblistes, reconnues pour leur efficacité sur les données tabulaires. Le troisième modèle introduit une approche connexionniste permettant d'exploiter une explication de type Grad-CAM. Le chapitre intègre également une logique de vote majoritaire afin de construire une décision finale plus robuste que celle d'un modèle unique.

Une attention particulière est accordée aux workflows. Dans un mémoire appliqué, la performance finale ne suffit pas : il faut aussi démontrer la cohérence de la chaîne complète, la reproductibilité du traitement, la séparation correcte entre apprentissage et test, ainsi que la traçabilité des explications. Pour cette raison, le chapitre est enrichi par des figures de pipeline, des tableaux de synthèse, des formules de métriques, des analyses critiques et une discussion méthodologique. La Fig. 4.1 donne une vue globale de la logique du chapitre.
Elle montre que l'implémentation ne se limite pas à l'entraînement des modèles, mais couvre toute la chaîne : préparation des données, modélisation, évaluation et interprétation.

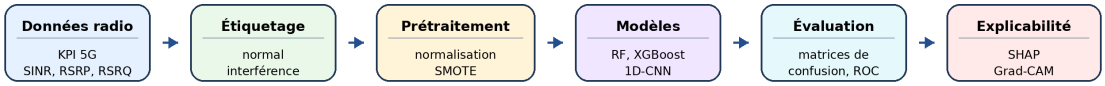


Fig. 4.1 – Pipeline de détection des interférences 5G par Machine Learning

Les contributions principales de ce chapitre peuvent être résumées comme suit. Premièrement, il formalise le pipeline expérimental sous une forme claire et reproductible. Deuxièmement, il détaille la logique d'étiquetage binaire et multi-classe utilisée dans le notebook. Troisièmement, il compare trois modèles sur un protocole commun. Quatrièmement, il introduit une double analyse d'explicabilité, à la fois par SHAP pour les modèles à arbres et par Grad-CAM pour le 1D-CNN. Enfin, il propose une lecture opérationnelle des résultats en vue d'une future intégration dans un système de supervision radio.

4.2 Objectifs expérimentaux du chapitre

Le chapitre poursuit quatre objectifs complémentaires. Le premier objectif est de vérifier que les KPI radio contiennent suffisamment d'information pour séparer les états normaux des états interférés. Les indicateurs utilisés sont notamment le SINR, le RSRP, le RSRQ, le RSSI, le débit descendant, le débit montant et le CQI. Ces grandeurs décrivent à la fois la puissance reçue, la qualité du canal et la capacité de transmission.

Le deuxième objectif est de comparer plusieurs approches d'apprentissage. Random Forest fournit une référence robuste, facile à entraîner et relativement interprétable. XGBoost apporte un mécanisme de boosting séquentiel capable de modéliser des frontières de décision très fines. Le 1D-CNN permet d'évaluer une architecture de deep learning compacte, même si les données sont tabulaires et non temporelles au sens strict.

Le troisième objectif est de construire une évaluation équilibrée. L'accuracy seule peut être trompeuse lorsqu'une classe est plus fréquente que l'autre. C'est pourquoi l'évaluation utilise aussi la précision, le rappel, le F1-score, l'AUC-ROC et les matrices de confusion. Ces métriques permettent de distinguer plusieurs comportements : détecter correctement les interférences, éviter les fausses alertes et produire des probabilités bien séparées.

Le quatrième objectif est d'expliquer les décisions. Dans le contexte des réseaux mobiles, un modèle qui déclenche une alerte sans justification exploitable est difficile à accepter par un ingénieur réseau. L'explicabilité permet de relier la décision automatique à des KPI physiques. Elle contribue donc à la confiance dans le système, mais aussi à la compréhension des causes possibles de la dégradation radio.

Tab. 4.1 – Objectifs expérimentaux associés au chapitre 4.

Objectif	Description	Sortie attendue
Détection	Identifier automatiquement les états radio interférés à partir des KPI.	Classe normal/interférence.
Comparaison	Évaluer Random Forest, XGBoost et 1D-CNN dans le même protocole.	Tableau de performances et courbes ROC.
Explicabilité	Identifier les variables qui influencent les décisions des modèles.	SHAP, importance native, Grad-CAM.
Opérationnalisation	Préparer le passage du notebook vers un système de monitoring.	Workflow de vote et pipeline temps réel.

4.3 Organisation générale du notebook

Le notebook utilisé comme base expérimentale suit une structure progressive. Les premières cellules installent et importent les bibliothèques nécessaires. Les cellules suivantes chargent ou génèrent le dataset, affichent les statistiques descriptives, créent les labels, visualisent les distributions et préparent les variables d'entrée. Les sections suivantes entraînent les modèles, calculent les métriques, génèrent les figures et appliquent les méthodes d'explicabilité.

Cette organisation présente un intérêt pédagogique important : chaque bloc du notebook correspond à une étape clairement identifiable du pipeline. En revanche, pour une intégration dans un mémoire, le notebook doit être transformé en un chapitre plus linéaire et plus argumenté. Le rôle de ce chapitre est donc de réécrire les sorties du notebook sous une forme scientifique, en explicitant les hypothèses, les limites et les interprétations.

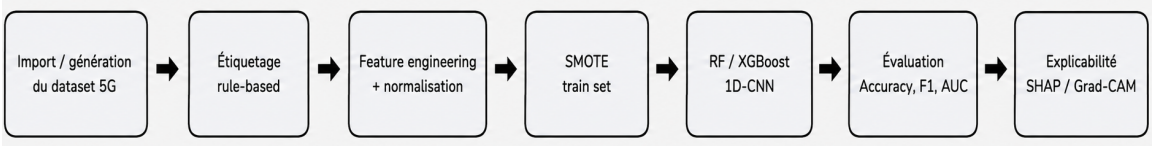


Fig. 4.2 – Chaîne expérimentale initiale générée à partir du notebook.

La Fig. 4.2 résume la chaîne initiale : les données sont transformées, les modèles sont entraînés, puis les décisions sont expliquées. La suite de ce chapitre détaille chacune de ces étapes.

4.4 Environnement de développement

4.4.1 Environnement logiciel

L'implémentation est réalisée en Python dans un environnement de type Google Colab. Ce choix facilite la reproductibilité pédagogique, car il évite une installation locale complexe et donne accès à un noyau Jupyter directement exécutable. Le notebook indique notamment une exécution avec TensorFlow 2.20.0 et sans GPU disponible. Dans cette configuration, les modèles à base d'arbres s'entraînent rapidement, tandis que le 1D-CNN reste suffisamment léger pour être entraîné sur CPU.

Tab. 4.2 – Stack logiciel utilisé pour l'implémentation.

Composant	Bibliothèque	Rôle dans l'expérimentation
Langage	Python	Langage hôte du pipeline complet de traitement et d'apprentissage.
Calcul numérique	NumPy	Manipulation vectorisée et génération du dataset synthétique.
Données tabulaires	Pandas	Lecture, organisation, nettoyage et transformation des KPI.
Visualisation	Matplotlib, Seaborn, Plotly	Figures de distribution, matrices de confusion, ROC, importance des variables.
Machine Learning	Scikit-learn	Split train/test, normalisation, Random Forest et métriques.
Boosting	XGBoost	Implémentation du modèle de gradient boosting sur données tabulaires.
Rééquilibrage	Imbalanced-learn	Application de SMOTE sur l'ensemble d'apprentissage uniquement.
Deep Learning	TensorFlow/Keras	Définition, entraînement et inférence du réseau 1D-CNN.
Explicabilité	SHAP	Analyse globale et locale des contributions de variables.
Explicabilité CNN	Grad-CAM	Visualisation des KPI activant le 1D-CNN.

Le choix de Python est cohérent avec les objectifs du projet. Les bibliothèques scientifiques permettent d'assembler rapidement un pipeline complet sans développer chaque algorithme depuis zéro. Scikit-learn fournit les composants standards du machine learning classique, XGBoost apporte une implémentation performante du boosting, TensorFlow/Keras permet de construire le modèle 1D-CNN, et SHAP offre une base solide pour l'explicabilité.

4.4.2 Contraintes d'exécution

L'environnement Colab est pratique, mais il introduit certaines contraintes. Les versions de bibliothèques peuvent évoluer avec le temps, les sessions sont temporaires, et le chemin d'accès aux fichiers dépend du montage de l'espace de travail. La reproductibilité impose donc de documenter les versions utilisées, de fixer les graines aléatoires lorsque cela est possible, et d'exporter les résultats importants sous forme de figures, tableaux CSV et fichiers LaTeX.

Dans l'exécution étudiée, le GPU n'est pas disponible. Cela ne pénalise pas fortement le projet, car le dataset comporte seulement 1000 observations et le 1D-CNN est compact. Pour des données réelles de grande taille, l'utilisation d'un GPU deviendrait plus importante, notamment pour les modèles séquentiels ou les architectures profondes plus complexes.

Tab. 4.3 – Contraintes pratiques observées et conséquences sur le travail.

Contrainte	Conséquence	Mesure recommandée
Session Colab temporaire	Les fichiers peuvent disparaître après déconnexion.	Sauvegarder notebook, figures et résultats dans une archive.
Versions variables	Les métriques peuvent légèrement varier selon les versions.	Documenter les versions et fixer les graines.
Absence du CSV réel	Génération automatique d'un dataset synthétique.	Remplacer par des mesures terrain pour la version finale.
Pas de GPU	Entraînement CNN plus lent mais acceptable ici.	GPU recommandé pour données plus volumineuses.
Données synthétiques	Généralisation terrain non démontrée.	Validation externe obligatoire.

4.5 Importation des données et contrôle initial

4.5.1 Logique d'importation

Le notebook cherche d'abord à lire un fichier CSV nommé `/content/B_2020.02.27_18.39.27.csv`. Lorsque ce fichier n'est pas présent dans l'environnement d'exécution, le notebook bascule vers une génération automatique d'un dataset de démonstration. Cette logique évite une interruption du notebook et permet de tester l'ensemble du pipeline même sans fichier externe.

La Fig. 4.3 représente cette logique. Elle est importante car elle explique pourquoi les résultats du chapitre ne doivent pas être interprétés comme des résultats obtenus sur la mesure réelle initialement prévue.

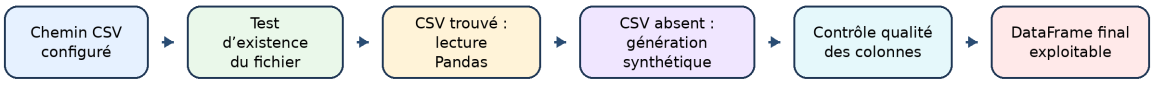


Fig. 4.3 – Workflow d'importation des données : lecture CSV ou génération synthétique.

Cette étape doit être conservée dans le code pour des raisons pédagogiques, mais elle doit être explicitement signalée dans le mémoire. Dans une version finale destinée à une soumission scientifique, il serait préférable de supprimer la génération automatique ou de l'encadrer

clairement comme un mode démonstration. L'expérience principale devrait alors reposer sur un fichier de mesures réel et archivé.

4.5.2 Profil statistique initial

Le dataset synthétique contient 1000 observations et huit colonnes initiales : SINR,RSRP, RSRQ, RSSI, débit descendant, débit montant, CQI et horodatage. Les statistiques descriptives indiquent des plages réalistes pour un exercice de simulation : le SINR varie approximativement de -4,88 dB à 19,99 dB, le RSRP de -119,84 dBm à -70,03 dBm, le RSRQ de -20 dB à -5,03 dB, et le RSSI de -109,97 dBm à -60,02 dBm.

Tab. 4.4 – Statistiques descriptives principales observées dans le notebook.

Variable	Moyenne	Minimum	Médiane	Maximum	Écart-type
SINR	7,26	-4,88	7,42	19,99	7,30
RSRP	-94,65	-119,84	-94,06	-70,03	14,61
RSRQ	-12,46	-20,00	-12,49	-5,03	4,36
RSSI	-85,48	-109,97	-85,79	-60,02	14,32
DL_bitrate	49410,53	3,07	49459,81	99774,94	28680,99
UL_bitrate	4984,64	61,84	4904,03	9993,50	2891,95
CQI	7,50	1,00	7,00	14,00	4,01

Ces statistiques permettent de vérifier que les variables couvrent plusieurs situations radio. Les valeurs négatives de SINR ou les RSRQ très dégradés représentent des conditions défavorables. Les valeurs élevées de SINR et les RSRQ moins négatifs représentent des conditions plus favorables. Cette diversité est nécessaire pour entraîner un classifieur binaire.

4.6 Description des KPI exploités

Le modèle utilise des KPI radio qui décrivent différentes facettes de la liaison 5G. Le SINR mesure le rapport entre le signal utile et le bruit. Le RSRP décrit la puissance reçue du signal de référence. Le RSRQ renseigne sur la qualité du signal de référence en tenant compte de la charge et des interférences. Le RSSI mesure la puissance totale reçue. Le CQI représente une indication de qualité de canal utilisée pour adapter la modulation et le codage. Les débits descendant et montant traduisent l'impact pratique de l'état radio sur le service.

Tab. 4.5 – Interprétation des KPI utilisés dans le chapitre.

KPI	Nature	Interprétation pour la détection d'interférence
SINR	Qualité signal/ interf.+bruit	Variable très discriminante : une baisse importante suggère une dégradation du lien.
RSRP	Puissance utile	Permet de distinguer une faible puissance reçue d'une dégradation liée au bruit ou aux interférences.
RSRQ	Qualité de référence	Indique la qualité globale du signal de référence ; très utile pour identifier une dégradation radio.
RSSI	Puissance totale reçue	Peut augmenter lorsque des sources interférentes contribuent à la puissance totale.
DL_bitrate	Service descendant	Mesure indirecte de l'effet de la dégradation radio sur le débit utile.
UL_bitrate	Service montant	Complète le débit descendant et reflète l'état du lien montant.
CQI	Qualité canal	Résume la capacité du canal à supporter des schémas de modulation/codage.
Timestamp	Temps	Conserve l'ordre des mesures, même si le modèle actuel ne l'exploite pas directement.

Il faut noter que le modèle actuel traite chaque ligne indépendamment. L'horodatage n'est pas utilisé comme variable d'entrée du classifieur. Cette décision simplifie le problème, mais elle empêche de modéliser la durée ou la récurrence d'un épisode d'interférence. Une perspective naturelle est donc d'introduire des fenêtres temporelles glissantes et des modèles séquentiels.

4.7 Étiquetage des données

4.7.1 Problématique de l'étiquetage

L'apprentissage supervisé exige une variable cible. Dans un contexte terrain, cette cible devrait idéalement provenir d'un diagnostic expert ou d'un système de mesure indépendant confirmant la présence d'une interférence. Dans le notebook, cette cible est construite automatiquement à partir de seuils sur les KPI. Cette approche est utile pour une preuve de concept, mais elle comporte une limite importante : le modèle peut apprendre à reproduire la règle qui a servi à créer les labels.

Cette limite n'annule pas l'intérêt de l'expérience. Elle signifie simplement que les résultats doivent être interprétés comme une validation technique du pipeline plutôt que comme une validation physique complète. Pour passer à une validation terrain, il faudrait disposer de labels indépendants, par exemple obtenus par mesures spectrales, logs opérateur, événements de brouillage identifiés ou annotations expertes.

4.7.2 Étiquetage binaire

La cible principale du notebook est binaire. Une observation est classée comme normale lorsque les KPI ne franchissent pas les seuils de dégradation définis. Elle est classée comme interférée lorsque le SINR est très faible ou lorsque le couple SINR-RSRQ indique une mauvaise qualité radio.

La règle finale appliquée peut s'écrire :

$$y = \begin{cases} 1, & \text{si SINR} < 0, \\ 1, & \text{si SINR} < 3 \text{ et } \mathrm{RSRQ} < -16, \\ 0, & \text{sinon.} \end{cases} \tag{4.1}$$

où $y = 1$ désigne une interférence et $y = 0$ désigne un état normal. La Fig. 4.4 présente cette logique sous forme de pipeline décisionnel.

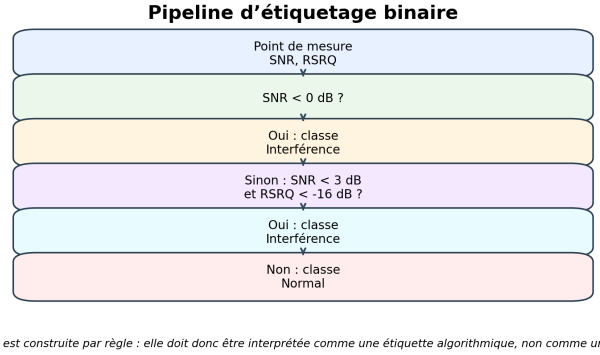


Fig. 4.4 – Pipeline d'étiquetage binaire utilisé pour construire la cible supervisée.

Après l'application de cette règle, le dataset contient 745 observations normales et 255 observations interférées. La classe interférence représente donc 25,5 % des observations. Ce déséquilibre est réel mais modéré. Il justifie l'utilisation de métriques complémentaires et l'application d'un rééquilibrage uniquement sur l'ensemble d'apprentissage.

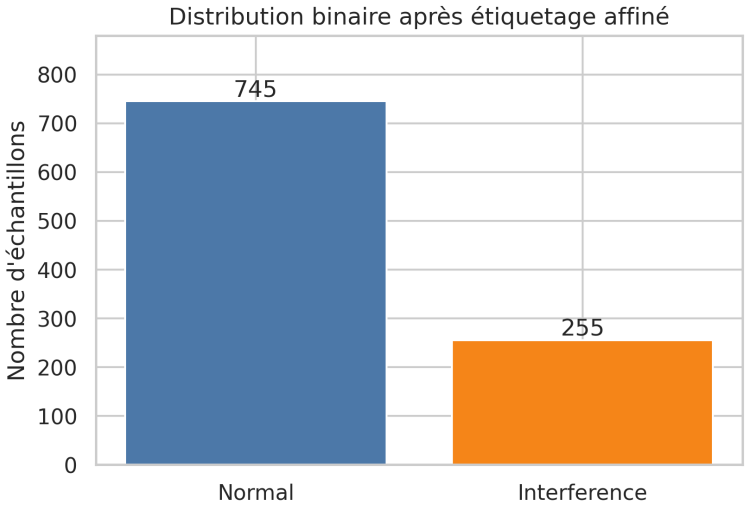


Fig. 4.5 – Distribution binaire du dataset après étiquetage affiné.

La Fig. 4.6 montre la séparation dans le plan SINR-RSRQ. Elle confirme visuellement que la règle d'étiquetage crée une frontière fortement dépendante de ces deux KPI.

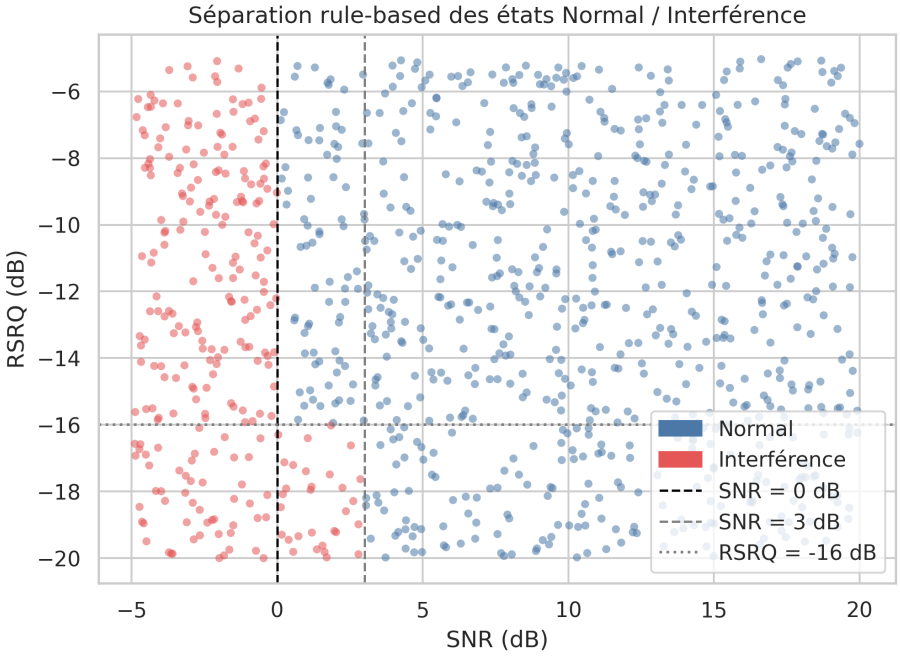


Fig. 4.6 – Visualisation de la règle d'étiquetage dans le plan SINR-RSRQ.

Cette figure est très utile pour comprendre les performances élevées des modèles d'arbres. Lorsque la cible est définie par des seuils simples, un arbre de décision ou une forêt d'arbres peut reconstruire très efficacement la règle. La performance élevée est donc cohérente, mais elle doit être accompagnée d'une discussion méthodologique.

4.7.3 Étiquetage multi-classe exploratoire

Le notebook propose également une extension multi-classe. Cette extension distingue quatre catégories : absence d'interférence, interférence co-canal, interférence de canal adjacent et interférence externe. Elle n'est pas utilisée comme cible principale des modèles comparés dans ce chapitre, mais elle donne une perspective intéressante pour la suite du travail.

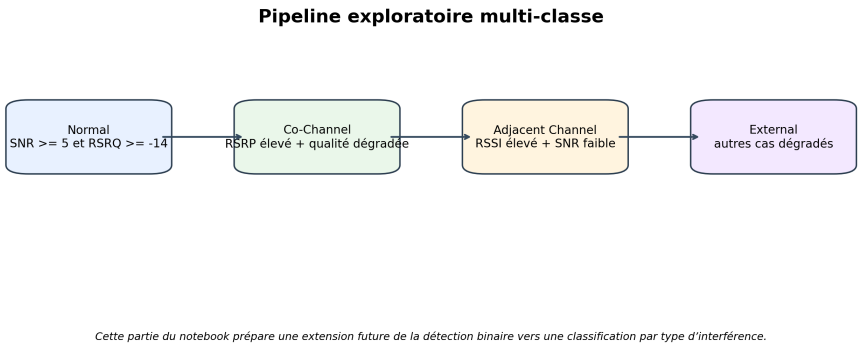


Fig. 4.7 – Pipeline exploratoire d'étiquetage multi-classe.

La répartition multi-classe obtenue est : 349 observations sans interférence, 349 observations de type externe, 219 observations co-canal et 83 observations de canal adjacent. Cette distribution est plus déséquilibrée que la distribution binaire, ce qui rendrait l'apprentissage multi-classe plus difficile.

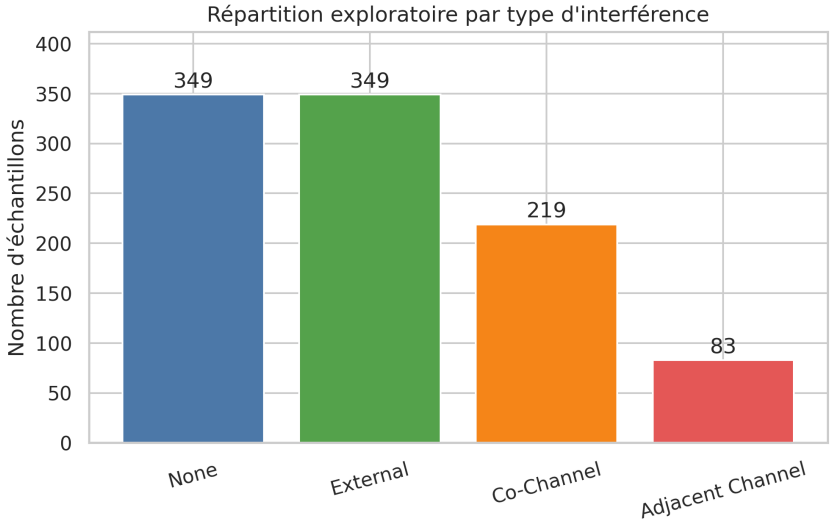


Fig. 4.8 – Répartition exploratoire des types d'interférences.

Tab. 4.6 – Règles exploratoires de classification multi-classe.

Classe	Condition indicative	Interprétation
None	SINR≥ 5 et $\mathrm{RSRQ} \geq -14$	Lien radio considéré comme stable.
Co-Channel	$\mathrm{RSRP} > -95$ et (SINR< 2 ou $\mathrm{RSRQ} < -17$)	Signal utile présent mais dégradation de qualité.
Adjacent Channel	$\mathrm{RSSI} > -85$ et SINR< 3	Puissance totale élevée avec faible qualité signal/bruit.
External	Autres cas de dégradation	Perturbation non expliquée par les règles précédentes.

Dans une étude future, cette classification devrait être validée par une expertise radio. Les seuils utilisés ici doivent être vus comme une première approximation, non comme une vérité universelle.

4.8 Feature engineering et préparation des variables

4.8.1 Vecteur d'entrée

Les modèles utilisent neuf variables d'entrée. Sept variables proviennent directement du dataset et deux variables sont créées par feature engineering. Le vecteur final peut être écrit :

$$\mathbf{x} = [\mathrm{SINR}, \mathrm{RSRP}, \mathrm{RSRQ}, DL, UL, CQI, RSSI, \mathrm{DL_UL_Ratio}, \mathrm{SINR_CQI_Product}]\,. \tag{4.2}$$

La variable $\mathrm{DL_UL_Ratio}$ compare le débit descendant et le débit montant :

$$\mathrm{DL_UL_Ratio} = \frac{\mathrm{DL_bitrate}}{\mathrm{UL_bitrate} + 10^{-6}}. \tag{4.3}$$

La variable $\mathrm{SINR_CQI_Product}$ combine le SINR et le CQI :

$$\mathrm{SINR_CQI_Product} = \mathrm{SINR} \times \mathrm{CQI}. \tag{4.4}$$

Cette deuxième variable est particulièrement informative, car elle amplifie l'effet d'un SINR dégradé lorsqu'il est associé à une faible qualité de canal. Elle apparaît d'ailleurs comme très importante dans l'analyse SHAP de XGBoost.

Tab. 4.7 – Variables finales utilisées par les modèles.

Variable	Origine	Rôle attendu
SINR	KPI brut	Détection directe d'une dégradation signal sur interférence+bruit.
RSRP	KPI brut	Information sur la puissance du signal de référence.
RSRQ	KPI brut	Qualité globale du signal de référence.
DL_bitrate	KPI brut	Impact descendant sur le service.
UL_bitrate	KPI brut	Impact montant sur le service.
CQI	KPI brut	Indicateur de qualité de canal.
RSSI	KPI brut	Puissance totale reçue.
DL_UL_Ratio	Variable dérivée	Déséquilibre entre trafic descendant et montant.
SINR_CQI_Product	Variable dérivée	Interaction entre qualité SINR et qualité canal.

4.8.2 Pipeline de prétraitement

La Fig. 4.9 présente le pipeline complet de prétraitement. Les variables sont d'abord sélectionnées, puis les deux variables dérivées sont calculées. Le dataset est ensuite divisé en un ensemble d'apprentissage et un ensemble de test selon une partition stratifiée 80/20. La normalisation est apprise uniquement sur l'ensemble d'apprentissage. Enfin, SMOTE est appliqué seulement à l'ensemble d'apprentissage.

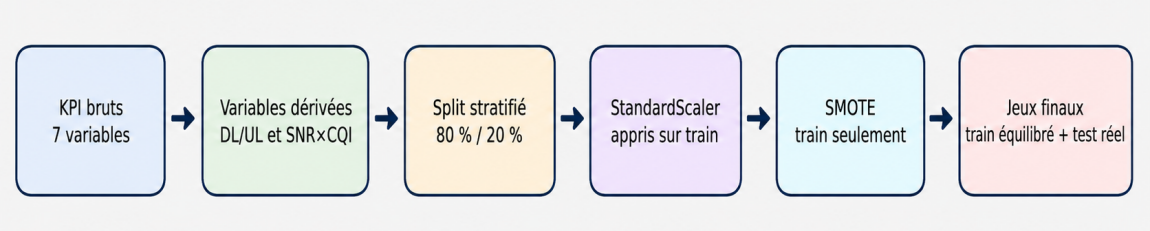


Fig. 4.9 – Pipeline de prétraitement et de feature engineering.

La normalisation est effectuée par score Z :

$$z_i = \frac{x_i - \mu_{\text{train}}}{\sigma_{\text{train}}}. \tag{4.5}$$

Les paramètres μ_{train} et σ_{train} sont calculés sur l'ensemble d'apprentissage. Ils sont ensuite appliqués au jeu de test. Cette règle évite la fuite d'information entre test et apprentissage.

4.8.3 Partition train/test et SMOTE

Le split stratifié conserve la proportion des classes. Le jeu de test contient 200 observations, dont environ 149 normales et 51 interférées. Le jeu d'apprentissage contient 800 observations

avant rééquilibrage. Après application de SMOTE, l'ensemble d'apprentissage atteint 1192 observations équilibrées, avec 596 exemples par classe.

Tab. 4.8 – Organisation des données après partition et rééquilibrage.

Ensemble	Taille	Classe normale	Classe interférence
Dataset complet	1000	745	255
Train avant SMOTE	800	596	204
Train après SMOTE	1192	596	596
Test final	200	149	51

Cette organisation respecte une règle essentielle : le jeu de test ne doit jamais être modifié par SMOTE. Il doit rester représentatif de la distribution que rencontrera le modèle en utilisation réelle. En revanche, l'ensemble d'apprentissage peut être rééquilibré pour aider les modèles à mieux apprendre la classe minoritaire.

4.9 Architecture expérimentale globale

4.9.1 Pipeline d'entraînement

Le pipeline d'entraînement est commun aux trois modèles afin de garantir une comparaison équitable. Chaque modèle reçoit les mêmes variables normalisées et le même ensemble d'apprentissage équilibré. Les différences de performance proviennent donc principalement de la capacité de modélisation de chaque algorithme.

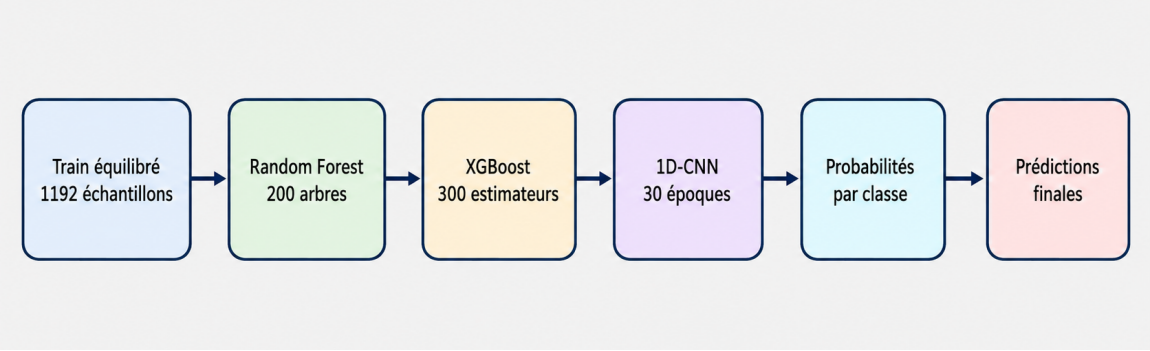


Fig. 4.10 – Workflow d'entraînement des trois modèles.

La sortie de chaque modèle comporte une prédiction discrète et, lorsque disponible, une probabilité d'appartenance à chaque classe. Les probabilités sont indispensables pour tracer les courbes ROC et pour fournir une confiance associée à la décision.

4.9.2 Pseudo-code général

Le fonctionnement global peut être résumé par le pseudo-code du Tableau 4.9. Ce pseudo-code reprend la logique principale du notebook sans dépendre d'une syntaxe Python détaillée.

Tab. 4.9 – Pseudo-code global du pipeline expérimental.

Étape	Instruction
1	Charger le fichier CSV si disponible ; sinon générer le dataset de démonstration.
2	Vérifier les colonnes KPI et afficher les statistiques descriptives.
3	Construire la cible binaire à partir des seuils sur SINR et RSRQ.
4	Construire éventuellement une cible multi-classe exploratoire.
5	Calculer les variables dérivées DL_UL_Ratio et SNR_CQI_Product.
6	Diviser les données en train/test avec stratification.
7	Normaliser les variables avec un StandardScaler appris sur le train.
8	Appliquer SMOTE uniquement sur le train normalisé.
9	Entraîner Random Forest, XGBoost et 1D-CNN.
10	Calculer les prédictions, les probabilités, les métriques et les figures.
11	Appliquer SHAP, Grad-CAM et comparer les importances des KPI.
12	Construire une décision finale par vote majoritaire pour les nouveaux points.

4.10 Modèle Random Forest

4.10.1 Principe

Random Forest est une méthode ensembliste fondée sur un ensemble d'arbres de décision. Chaque arbre est entraîné sur un sous-échantillon bootstrap de l'ensemble d'apprentissage et utilise une sélection partielle de variables lors des divisions. La prédiction finale est obtenue par vote majoritaire. Ce mécanisme réduit la variance et limite le sur-apprentissage d'un arbre unique.

Dans ce projet, Random Forest est particulièrement adapté car le problème repose sur des seuils et interactions entre KPI. Les arbres de décision sont naturellement capables de modéliser des règles du type : si le SINR est inférieur à une valeur et si le RSRQ est inférieur à une autre valeur, alors la probabilité d'interférence augmente.

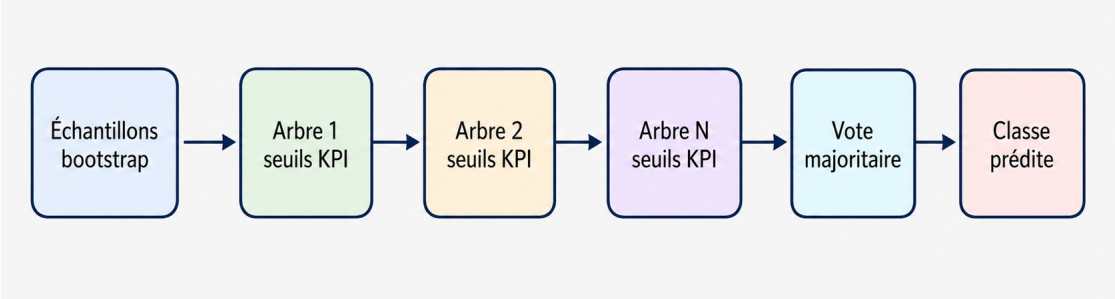


Fig. 4.11 – Principe simplifié du Random Forest appliqué aux KPI radio.

4.10.2 Paramètres utilisés

Le notebook utilise 200 arbres et une profondeur maximale de 12. Cette configuration représente un compromis entre capacité de modélisation et contrôle de la complexité. Une profondeur trop faible pourrait sous-apprendre la règle, tandis qu'une profondeur trop élevée pourrait mémoriser des particularités du dataset synthétique.

Tab. 4.10 – Hyperparamètres du Random Forest.

Paramètre	Valeur
Nombre d'arbres	200
Profondeur maximale	12
Graine aléatoire	42
Parallélisation	Tous les coeurs CPU disponibles
Entrée	9 variables normalisées
Sortie	Classe binaire et probabilité d'interférence

L'autre avantage du Random Forest est son importance native des variables. Même si cette importance doit être interprétée avec prudence, elle fournit un premier niveau d'explication du modèle.

4.11 Modèle XGBoost

4.11.1 Principe

XGBoost repose sur le gradient boosting. Contrairement au Random Forest, les arbres ne sont pas entraînés indépendamment. Ils sont ajoutés progressivement, chaque nouvel arbre cherchant à corriger les erreurs résiduelles de l'ensemble déjà construit. Cette approche permet de produire des modèles très performants sur les données tabulaires.

La Fig. 4.12 illustre ce mécanisme. Le modèle commence par une prédiction initiale, calcule les erreurs, puis ajoute successivement des arbres correcteurs. La prédiction finale est la somme pondérée des contributions de tous les arbres.

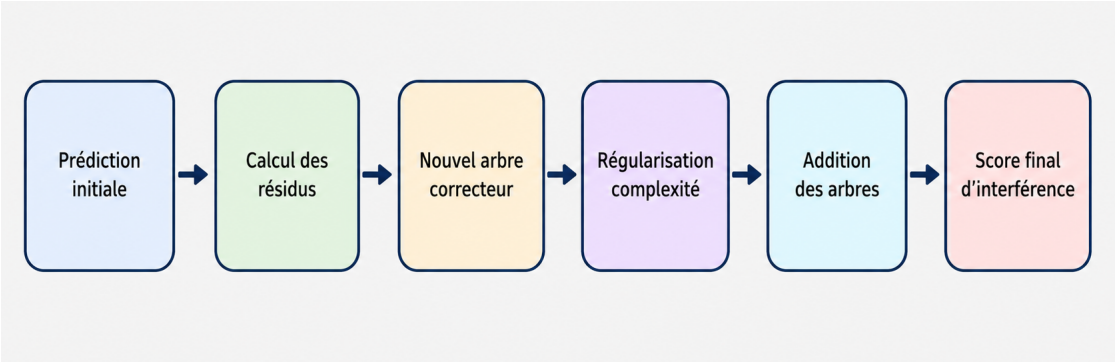


Fig. 4.12 – Principe simplifié de XGBoost : correction séquentielle des erreurs.

4.11.2 Paramètres utilisés

Les paramètres retenus dans le notebook sont 300 estimateurs, un taux d'apprentissage de 0,05 et une profondeur maximale de 8. Le taux d'apprentissage faible rend l'ajout de chaque arbre plus progressif. Le nombre d'estimateurs plus élevé compense cette progression lente.

Tab. 4.11 – Hyperparamètres du modèle XGBoost.

Paramètre	Valeur
Nombre d'estimateurs	300
Taux d'apprentissage	0,05
Profondeur maximale	8
Graine aléatoire	42
Entrée	9 variables normalisées
Sortie	Classe binaire et probabilité d'interférence

XGBoost est aussi très utile pour l'explicabilité, car il est compatible avec SHAP. Les valeurs SHAP permettent d'analyser les contributions de chaque KPI à la probabilité d'interférence.

4.12 Modèle 1D-CNN

4.12.1 Justification du 1D-CNN

Le 1D-CNN est introduit pour tester une approche de deep learning et pour permettre une explication de type Grad-CAM. Cependant, il faut être précis : les données utilisées ici sont

tabulaires. Le vecteur de neuf KPI est réorganisé sous la forme d'une séquence artificielle de longueur 9. La convolution ne porte donc pas sur un signal temporel continu, mais sur un ordre fixe de variables.

Cette utilisation reste possible à titre expérimental, mais elle doit être interprétée avec prudence. Si l'ordre des variables est modifié, les motifs locaux appris par le CNN peuvent changer. Pour une version terrain plus solide, il serait préférable d'utiliser des fenêtres temporelles de KPI ou des modèles séquentiels, par exemple LSTM, GRU ou Transformer.

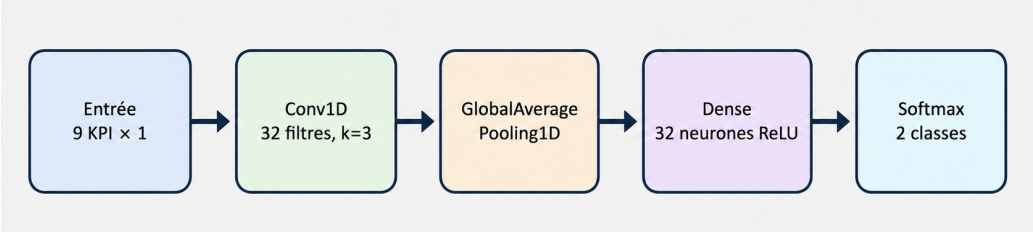


Fig. 4.13 – Architecture 1D-CNN utilisée pour traiter le vecteur des KPI.

4.12.2 Architecture et entraînement

L'entrée du modèle est de taille 9×1. Elle passe par une couche Conv1D de 32 filtres avec noyau de taille 3, puis par un GlobalAveragePooling1D. Une couche dense de 32 neurones avec activation ReLU précède la couche finale Softmax à deux classes.

Tab. 4.12 – Architecture du 1D-CNN.

Bloc	Paramètre	Rôle
Entrée	9×1	Représentation des neuf KPI normalisés.
Conv1D	32 filtres, noyau 3	Extraction de motifs locaux entre KPI voisins.
AveragePooling1D	Moyenne globale	Réduction de la dimension avant classification.
Dense	32 neurones ReLU	Combinaison non linéaire des représentations.
Sortie	Softmax 2 classes	Probabilités normal/interférence.

Le modèle est entraîné pendant 30 époques avec un batch de 64 exemples. La fonction de perte est l'entropie croisée catégorielle creuse et l'optimiseur utilisé est Adam. Les courbes d'apprentissage sont présentées dans la Fig. 4.14.

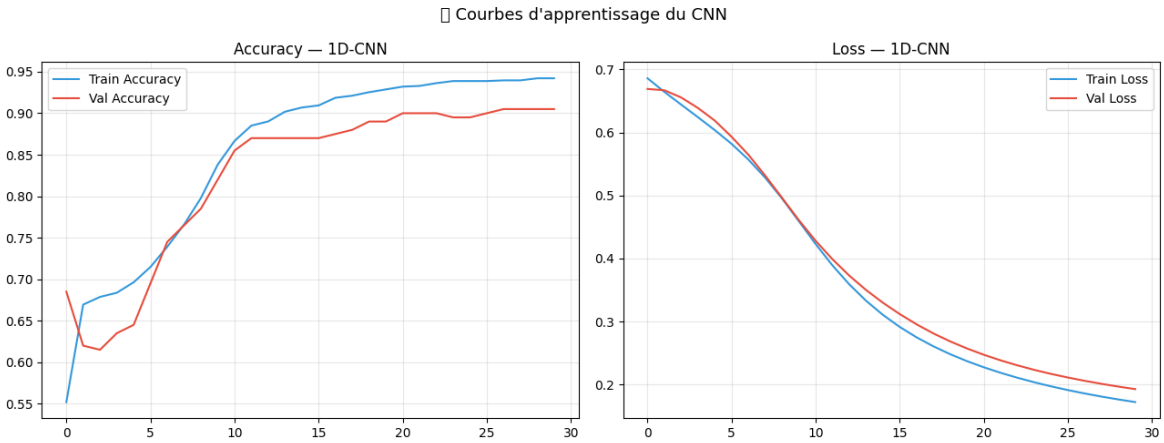


Fig. 4.14 – Courbes d'apprentissage du 1D-CNN sur 30 époques.

Ces courbes permettent de vérifier la convergence globale du modèle. Dans l'exécution étudiée, le 1D-CNN atteint une accuracy inférieure à celle des modèles d'arbres, mais il conserve une AUC élevée. Cette observation suggère que la séparation probabiliste est correcte, même si le seuil de décision par défaut n'est pas optimal.

4.13 Vote majoritaire et décision finale

Le notebook propose une logique de vote majoritaire pour combiner les trois modèles. Le principe est simple : chaque modèle produit une classe, puis la décision finale correspond à la classe obtenant au moins deux votes sur trois. Cette stratégie peut améliorer la robustesse lorsque les modèles possèdent des erreurs partiellement différentes.

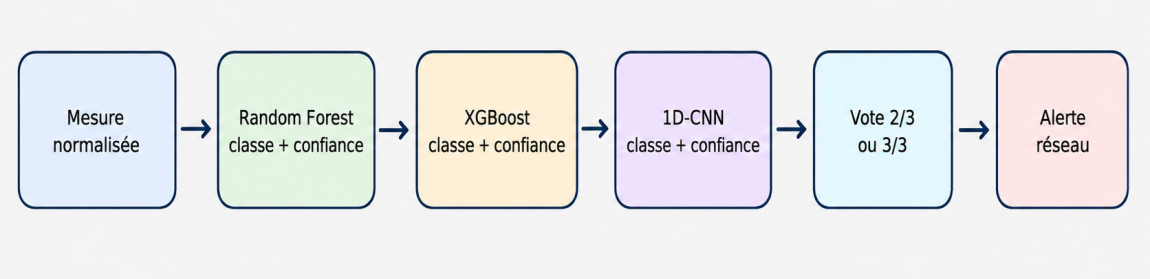


Fig. 4.15 – Pipeline de décision par vote majoritaire entre Random Forest, XGBoost et 1D-CNN.

Le vote majoritaire présente deux avantages. D'abord, il réduit la dépendance à un modèle unique. Ensuite, il facilite l'interprétation opérationnelle : si les trois modèles sont d'accord, l'alerte est plus crédible ; si un seul modèle détecte une interférence, la décision peut être considérée comme incertaine. Une amélioration future consisterait à pondérer les votes par la confiance ou par la performance validée de chaque modèle.

Tab. 4.13 – Exemples d'interprétation du vote majoritaire.

Random Forest	XGBoost	1D-CNN	Décision finale
Normal	Normal	Normal	Normal, confiance forte.
Interférence	Interférence	Interférence	Alerte, confiance forte.
Interférence	Interférence	Normal	Alerte, confiance moyenne à forte.
Normal	Interférence	Normal	Pas d'alerte, mais cas à surveiller.
Interférence	Normal	Normal	Pas d'alerte, décision incertaine.

4.14 Métriques d'évaluation

4.14.1 Matrice de confusion

Pour un problème binaire, la matrice de confusion compare la classe réelle et la classe prédite. Les quatre quantités fondamentales sont les vrais positifs (TP), vrais négatifs (TN), faux positifs (FP) et faux négatifs (FN). Dans ce chapitre, la classe positive correspond à l'état interférence.

Tab. 4.14 – Lecture de la matrice de confusion pour la détection d'interférence.

Quantité	Cas	Signification opérationnelle
TP	Interférence réelle prédite interférence	Détection correcte d'un événement perturbateur.
TN	Normal réel prédit normal	Absence correcte d'alerte.
FP	Normal réel prédit interférence	Fausse alerte pouvant déclencher une intervention inutile.
FN	Interférence réelle prédite normal	Interférence non détectée, potentiellement plus critique.

4.14.2 Formules des métriques

L'accuracy mesure la proportion globale de prédictions correctes :

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.6}$$

La précision mesure la proportion d'alertes réellement justifiées :

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{4.7}$$

Le rappel mesure la proportion d'interférences réellement détectées :

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{4.8}$$

Le F1-score est la moyenne harmonique de la précision et du rappel :

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4.9}$$

L'AUC-ROC mesure la capacité du modèle à séparer les deux classes lorsque le seuil de décision varie. Elle complète le F1-score, car un modèle peut avoir une bonne AUC mais un F1-score inférieur si le seuil choisi n'est pas optimal.

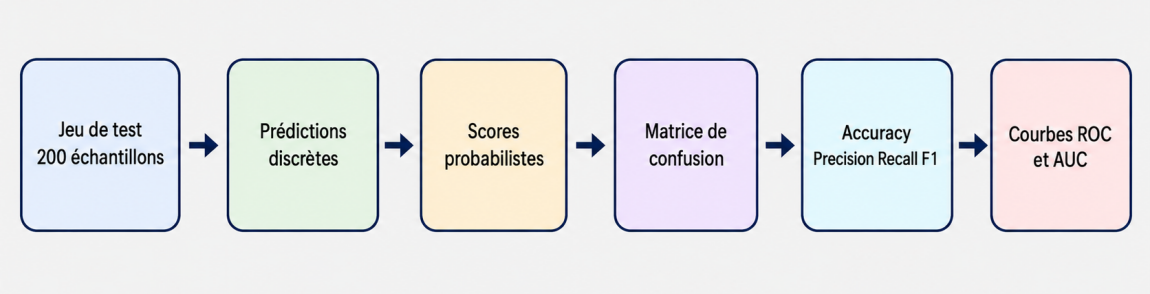


Fig. 4.16 – Pipeline d'évaluation : prédictions, probabilités, matrices et métriques.

4.15 Résultats expérimentaux

4.15.1 Synthèse quantitative

Le Tableau 4.15 présente les résultats obtenus sur le jeu de test de 200 échantillons. Les modèles à base d'arbres dominent clairement le 1D-CNN sur l'accuracy et le F1-score.

Tab. 4.15 – Performances des modèles sur le jeu de test.

Modèle	Accuracy	Précision	Rappel	F1-score	AUC-ROC
Random Forest	100,00 %	100,00 %	100,00 %	1,0000	1,0000
XGBoost	99,50 %	98,08 %	100,00 %	0,9903	1,0000
1D-CNN	90,50 %	75,00 %	94,12 %	0,8348	0,9833

Ces résultats doivent être lus en deux niveaux. Au niveau technique, ils montrent que le pipeline fonctionne correctement et que les modèles apprennent la séparation imposée par

les règles d'étiquetage. Au niveau scientifique, ils doivent être interprétés avec prudence, car la cible est fortement liée au SINR et au RSRQ, qui sont aussi fournis en entrée aux modèles.

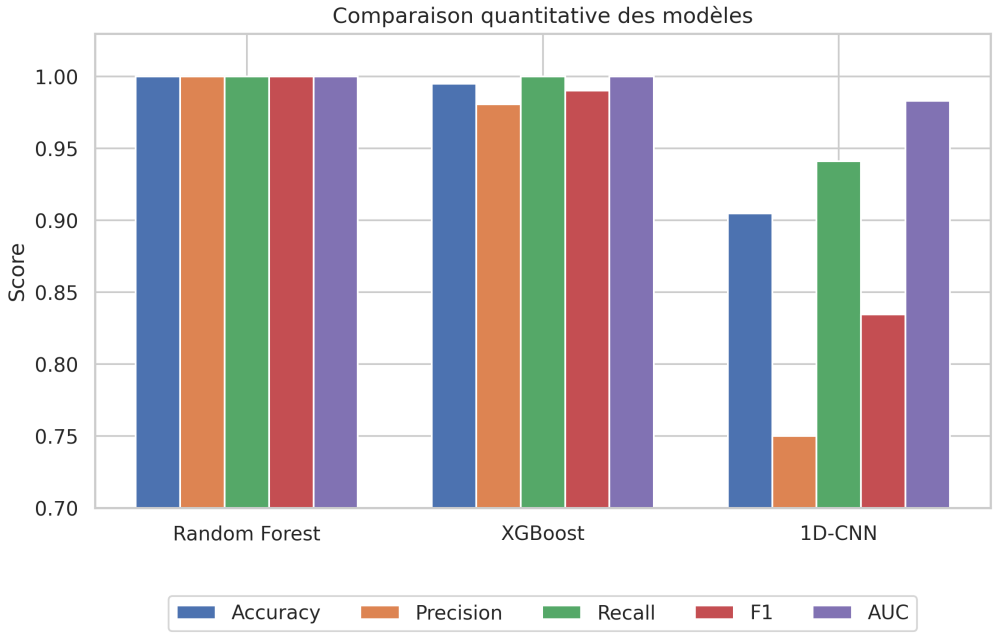


Fig. 4.17 – Comparaison graphique des métriques de performance.

4.15.2 Analyse des matrices de confusion

La Fig. 4.18 présente les matrices de confusion. Le Random Forest classe correctement tous les échantillons du test. XGBoost produit une seule fausse alerte. Le 1D-CNN détecte la majorité des interférences, mais il génère davantage de fausses alertes.

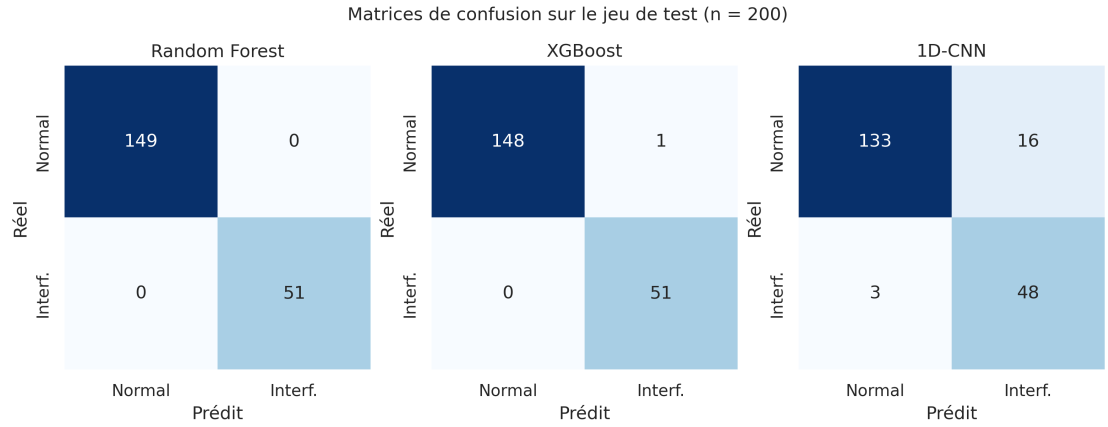


Fig. 4.18 – Matrices de confusion obtenues sur le jeu de test.

Cette différence a une signification opérationnelle. Un faux positif déclenche une alerte inutile. Un faux négatif laisse une interférence non détectée. Dans un réseau réel, le compromis entre ces deux erreurs dépend du coût d'intervention et du niveau de criticité du service. Pour

des services critiques, le rappel peut être privilégié. Pour des réseaux fortement supervisés, une précision élevée peut être préférée afin de limiter les alarmes inutiles.

4.15.3 Courbes ROC

Les courbes ROC montrent la sensibilité du modèle au seuil de décision. Le Random Forest et XGBoost atteignent une AUC égale à 1, tandis que le 1D-CNN atteint 0,9833. Le résultat du 1D-CNN est donc meilleur que ne le laisse supposer son F1-score. Cela signifie que ses scores probabilistes contiennent une information utile, mais que le seuil de décision par défaut n'est pas optimal.

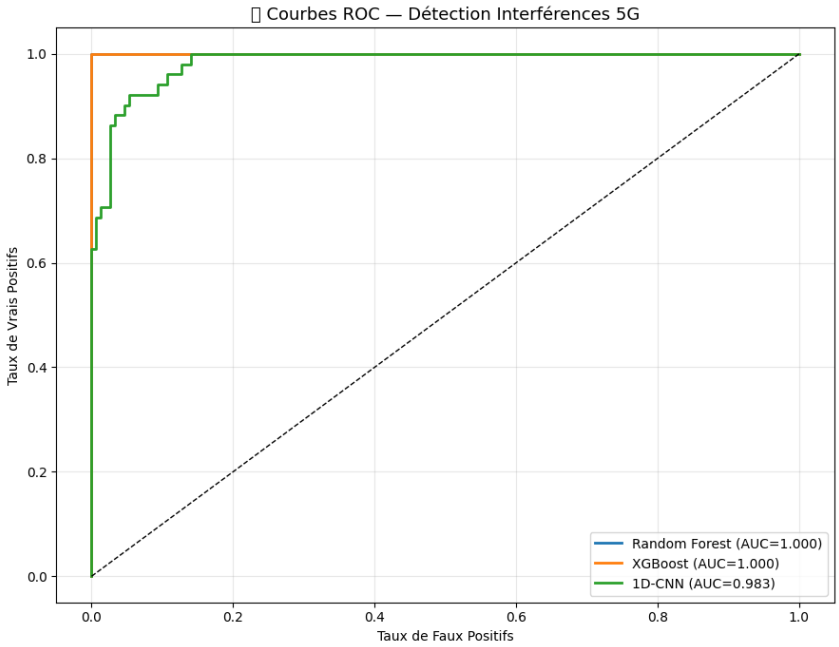


Fig. 4.19 – Courbes ROC des trois modèles générées dans le notebook.

La Fig. 4.20 présente également une version focalisée sur les modèles d'arbres. Ces courbes confirment que les deux modèles séparent presque parfaitement les classes dans ce dataset synthétique.

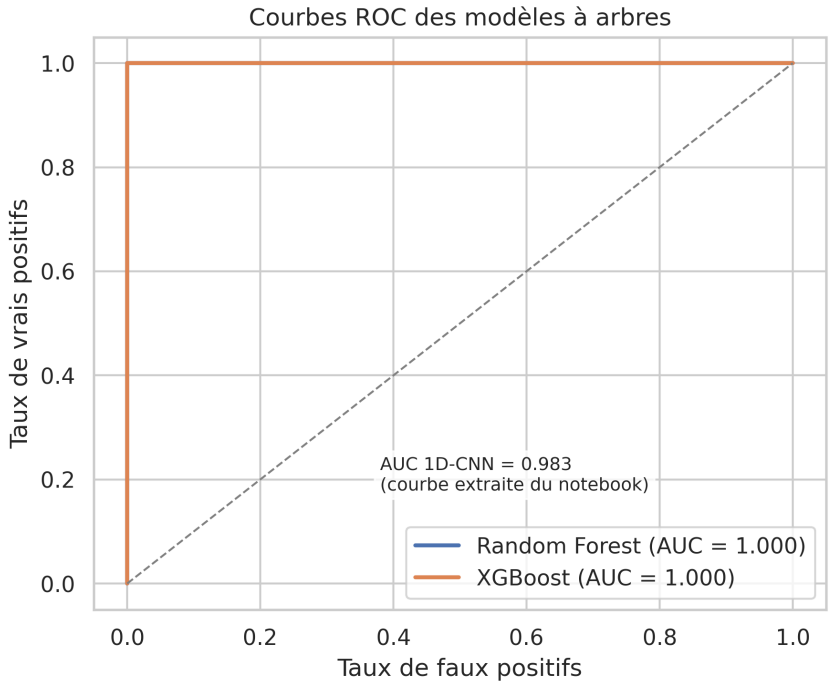


Fig. 4.20 – Courbes ROC focalisées sur Random Forest et XGBoost.

4.15.4 Relation entre SINR et probabilité d'interférence

La Fig. 4.21 montre la relation entre le SINR et la probabilité prédite d'interférence. Cette figure est importante pour vérifier la cohérence physique du modèle. Lorsque le SINR diminue fortement, la probabilité d'interférence augmente. Cette tendance correspond à l'intuition radio.

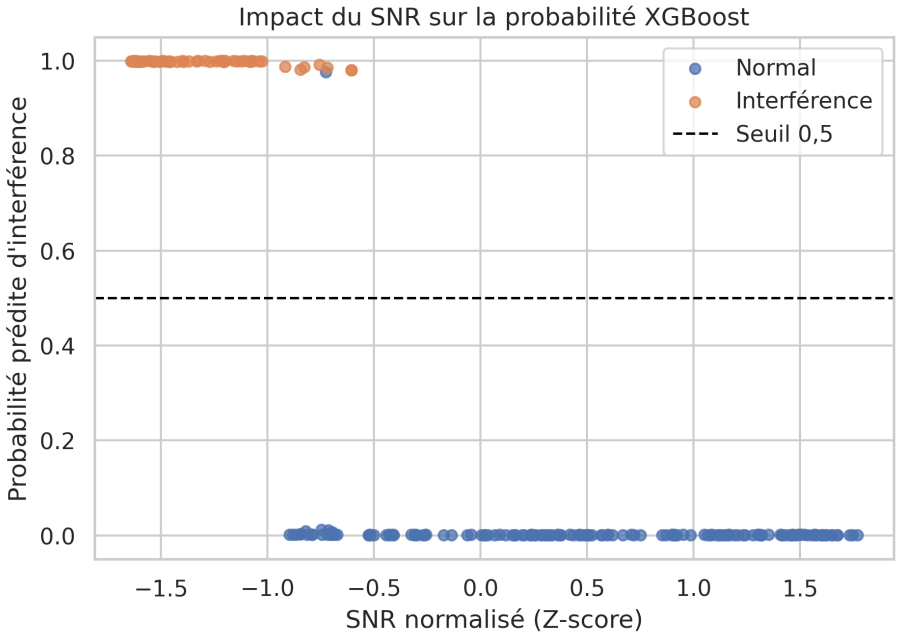


Fig. 4.21 – Relation entre SINR et probabilité d'interférence prédite.

Cependant, cette cohérence doit être distinguée d'une validation indépendante. Le SINR est aussi utilisé pour construire les labels. Le modèle est donc cohérent avec la règle d'étiquetage, ce qui est attendu. Une validation plus forte consisterait à vérifier que la probabilité prédite reste pertinente sur des données annotées par un expert et non par la même règle.

4.16 Explicabilité par SHAP

4.16.1 Principe général

SHAP attribue à chaque variable une contribution à la prédiction. Pour un modèle donné et un échantillon donné, les valeurs SHAP indiquent quelles variables poussent la prédiction vers la classe interférence et quelles variables la poussent vers la classe normale. L'intérêt de SHAP est de fournir à la fois une explication globale et une explication locale.

Dans ce travail, SHAP est appliqué principalement à XGBoost. Ce choix est pertinent car les modèles d'arbres sont bien supportés par les algorithmes SHAP rapides et stables. L'objectif est de vérifier que le modèle ne s'appuie pas sur des variables arbitraires, mais sur des KPI cohérents avec le phénomène étudié.

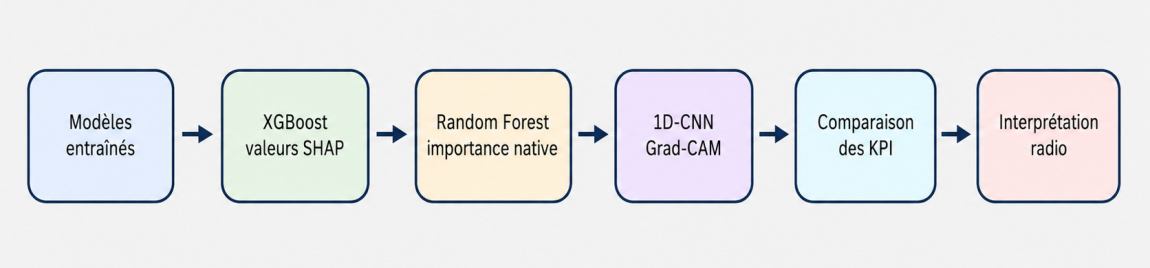


Fig. 4.22 – Workflow d'explicabilité combinant SHAP, importance native et Grad-CAM.

4.16.2 Importance globale des variables

La Fig. 4.23 montre l'importance globale des variables selon SHAP. Les variables les plus influentes sont SNR_CQI_Product, SINRet RSRQ. Cette hiérarchie est cohérente avec la règle d'étiquetage et avec l'interprétation physique du problème.

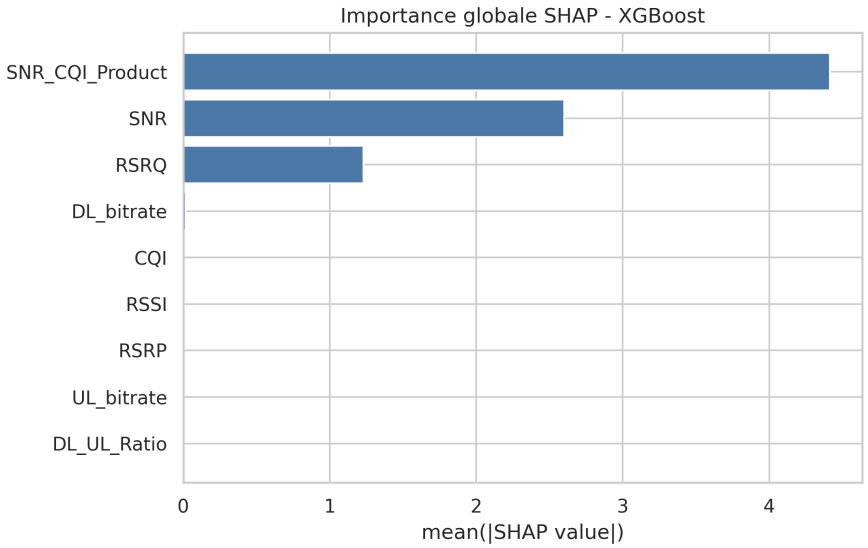


Fig. 4.23 – Importance globale des variables selon SHAP pour XGBoost.

Le fait que SNR_CQI_Product domine le classement montre l'intérêt du feature engineering. Cette variable combine directement une information de qualité signal/bruit et une information de qualité de canal. Elle peut donc résumer une partie de l'état radio sous une forme plus exploitable par le modèle.

4.16.3 Diagramme beeswarm

Le diagramme beeswarm de la Fig. 4.24 donne une lecture plus détaillée. Chaque point représente une observation. La position horizontale indique l'effet de la variable sur la prédiction, tandis que la couleur représente la valeur de la variable.

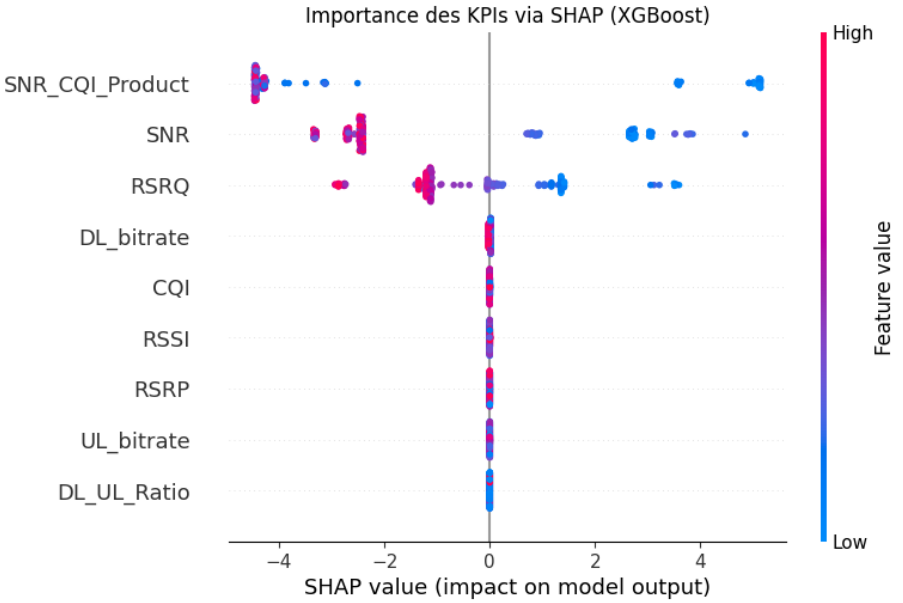


Fig. 4.24 – Diagramme SHAP beeswarm pour le modèle XGBoost.

Une valeur SHAP positive pousse la prédiction vers la classe interférence. Une valeur négative la pousse vers la classe normale. Cette visualisation permet donc d'identifier non seulement les variables importantes, mais aussi le sens de leur effet. Elle est particulièrement utile dans un contexte d'aide au diagnostic.

4.16.4 Explication locale et analyse d'erreur

Le notebook contient aussi des explications locales, notamment pour des cas mal classés. Cette partie est importante car un système de monitoring doit être capable d'expliquer pourquoi une alerte a été déclenchée. Une explication locale peut indiquer, par exemple, que la décision est surtout due à un SINRfaible ou à une combinaison défavorable entre SINRet CQI.

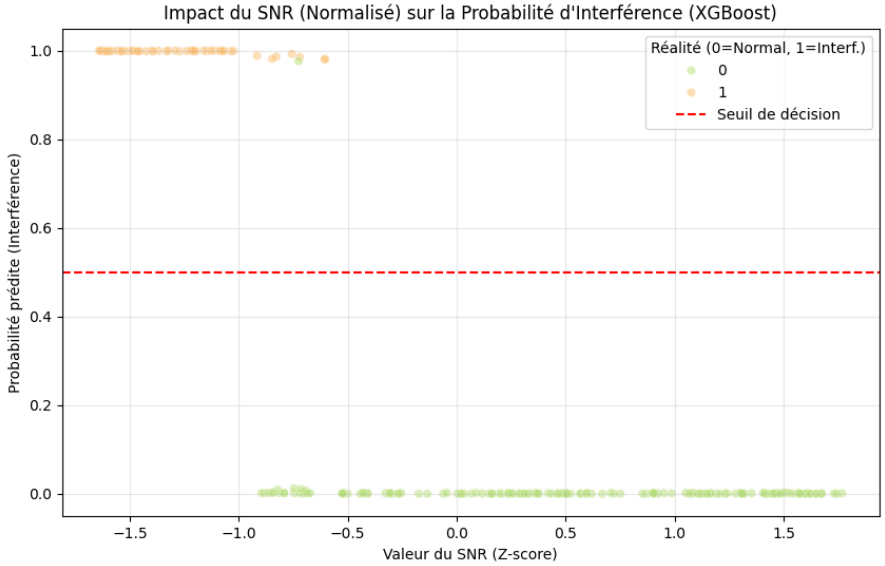


Fig. 4.25 – Exemple d'analyse locale issue du notebook.

Lorsque des erreurs apparaissent, l'analyse locale aide à distinguer deux situations. Dans la première, l'erreur provient d'une frontière de décision trop stricte ou trop souple. Dans la seconde, l'erreur peut provenir d'une incohérence de label, surtout si la cible a été générée par règle. Cette distinction est essentielle pour améliorer le système.

4.17 Explicabilité par Grad-CAM

4.17.1 Principe appliqué au 1D-CNN

Grad-CAM est généralement utilisé pour expliquer les réseaux convolutionnels appliqués aux images. Dans ce projet, il est adapté au 1D-CNN afin de visualiser les zones du vecteur

d'entrée qui activent le plus la couche convolutionnelle. Chaque position du vecteur correspond à un KPI. Le Grad-CAM indique donc quels KPI ont le plus contribué à la décision du CNN.

Il faut rappeler que cette interprétation dépend de l'ordre des variables dans le vecteur. Contrairement à une image, où la proximité spatiale est naturelle, la proximité entre deux KPI est ici définie artificiellement par l'ordre choisi dans le tableau de variables. Cette limite ne rend pas l'analyse inutile, mais elle impose une interprétation prudente.

4.17.2 Grad-CAM global

Le notebook calcule un Grad-CAM moyen sur les cas d'interférence du jeu de test. Le SINR apparaît comme le KPI le plus activant, suivi de SNR_CQI_Product, du débit montant et du RSRP.

Tab. 4.16 – Classement des KPI selon le Grad-CAM moyen du 1D-CNN.

KPI	Score moyen	Écart-type
SINR	0,641	0,413
SNR_CQI_Product	0,351	0,415
UL_bitrate	0,339	0,422
RSRP	0,286	0,333
DL_bitrate	0,250	0,361
DL_UL_Ratio	0,188	0,286
CQI	0,156	0,219
RSSI	0,108	0,181
RSRQ	0,089	0,170

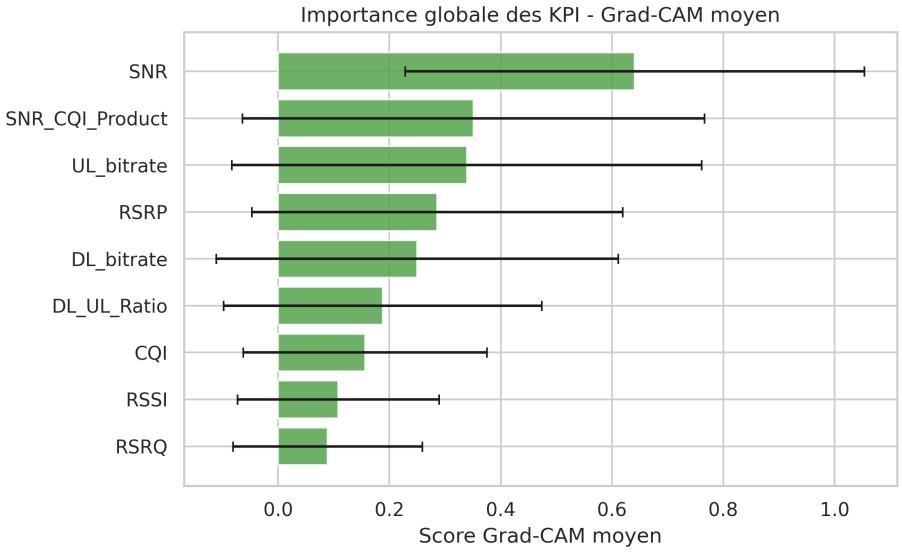


Fig. 4.26 – Importance globale des KPI selon le Grad-CAM moyen du 1D-CNN.

La dominance du SINR est cohérente avec le mécanisme d'étiquetage. En revanche, le score relativement faible du RSRQ dans le Grad-CAM montre que le CNN n'exploite pas exactement les variables de la même manière que XGBoost. Cette différence peut provenir de l'architecture, de l'ordre des variables ou du seuil de décision.

4.17.3 Grad-CAM sur un nouveau cas

Le notebook applique aussi le modèle à un nouveau point de mesure. Les trois modèles prédisent une interférence. Le vote majoritaire déclenche donc une alerte. L'explication Grad-CAM du nouveau cas indique que les trois KPI les plus déterminants sont SNR_CQI_Product, CQI et RSSI.

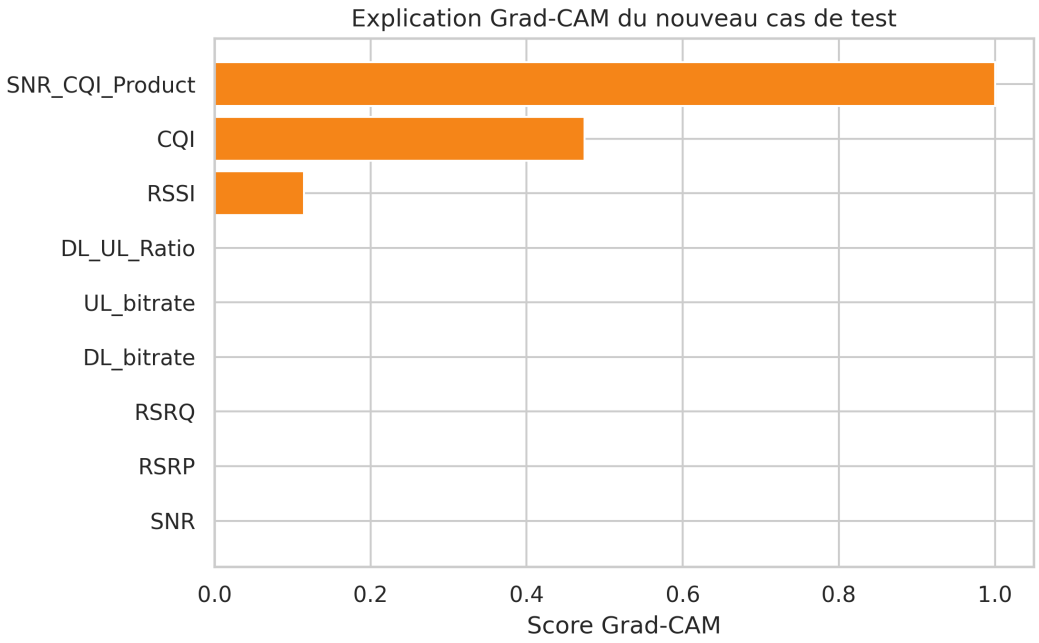


Fig. 4.27 – Explication Grad-CAM pour un nouveau point de mesure classé interférence.

Cet exemple illustre l'intérêt opérationnel de l'explicabilité. Le système ne se contente pas d'indiquer qu'une interférence est détectée ; il peut aussi fournir les KPI qui ont motivé l'alerte. Cette information peut orienter l'ingénieur réseau vers une investigation plus rapide.

4.18 Comparaison croisée des importances

Les trois méthodes d'explicabilité ne mesurent pas exactement la même quantité. L'importance native du Random Forest mesure la contribution moyenne d'une variable à la réduction d'impureté dans les arbres. SHAP mesure une contribution marginale à la prédiction. Grad-CAM mesure l'activation d'une couche convolutionnelle par les positions d'entrée.

Malgré ces différences, la comparaison est utile. Elle permet de vérifier si plusieurs modèles, entraînés différemment, convergent vers les mêmes KPI. La Fig. 4.28 montre cette comparaison.

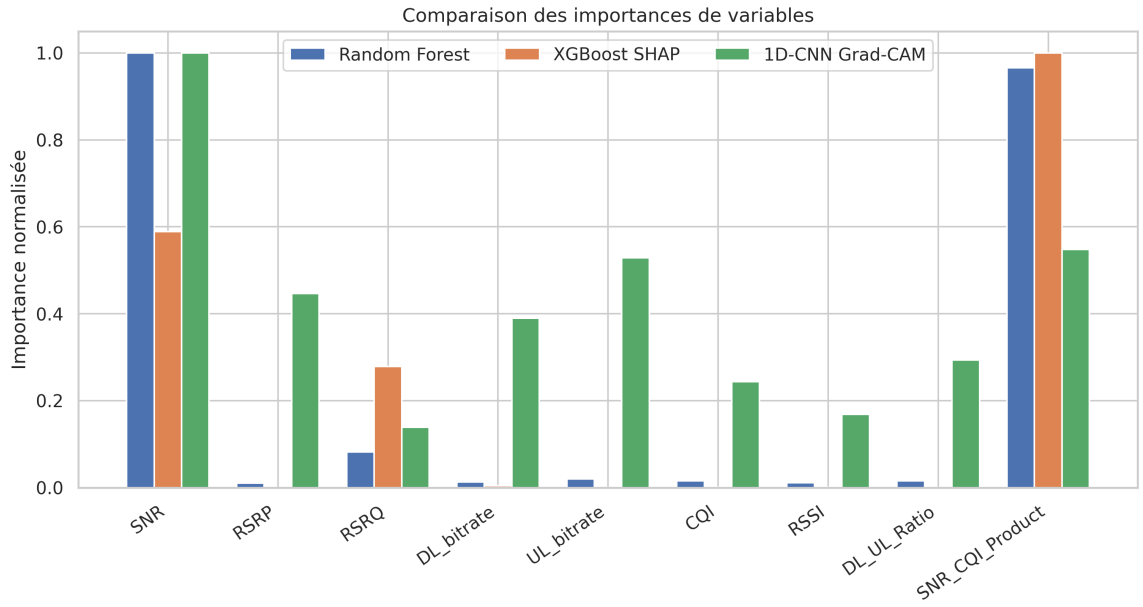


Fig. 4.28 – Comparaison des importances normalisées : Random Forest, SHAP et Grad-CAM.

Les résultats confirment l'importance du SINR et de la variable SNR_CQI_Product. Le RSRQ est très important pour les modèles à arbres, mais moins dominant pour le 1D-CNN selon Grad-CAM. Cette divergence est intéressante car elle montre que le deep learning ne reproduit pas exactement la logique des arbres.

Tab. 4.17 – Lecture comparative des méthodes d'explicabilité.

Méthode	Modèle concerné	Interprétation
Importance native	Random Forest	Mesure liée à la réduction d'impureté dans les arbres.
SHAP	XGBoost	Contribution moyenne et locale des variables à la prédiction.
Grad-CAM	1D-CNN	Activation des positions du vecteur d'entrée dans la couche convolutionnelle.
Comparaison croisée	Tous	Vérification de la cohérence globale entre plusieurs familles de modèles.

4.19 Dashboard et monitoring opérationnel

4.19.1 Passage du notebook au monitoring

Le notebook contient une logique de diagnostic sur de nouveaux points de mesure. Cette logique peut être vue comme un prototype de dashboard de monitoring. En pratique, un flux de KPI serait collecté depuis le réseau, normalisé avec les mêmes paramètres que l'entraînement, puis envoyé aux trois modèles. La décision finale serait construite par vote majoritaire et enrichie par une explication.

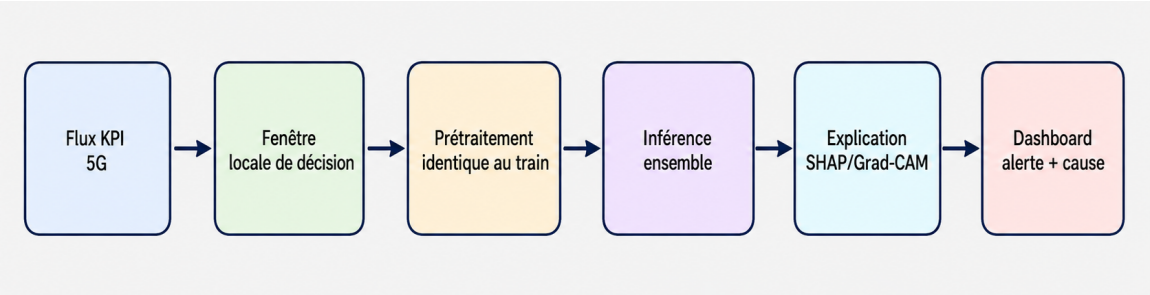


Fig. 4.29 – Pipeline proposé pour un monitoring temps réel des interférences.

Le monitoring temps réel impose des contraintes supplémentaires. Le modèle doit répondre rapidement, les données manquantes doivent être gérées, les distributions doivent être surveillées, et les seuils d'alerte doivent être adaptés au coût opérationnel des fausses alertes.

4.19.2 Indicateurs à afficher dans un tableau de bord

Un tableau de bord opérationnel ne doit pas afficher uniquement une classe. Il doit fournir un ensemble d'informations permettant à un ingénieur de comprendre et de prioriser les alertes. Les indicateurs recommandés sont donnés dans le Tableau 4.18.

Tab. 4.18 – Indicateurs recommandés pour un tableau de bord opérationnel.

Indicateur	Utilité	Priorité
Classe prédite	Indique normal ou interférence.	Très élevée
Probabilité d'interférence	Mesure la confiance du modèle.	Très élevée
Vote des trois modèles	Permet de distinguer accord fort et décision incertaine.	Élevée
Top KPI explicatifs	Oriente l'analyse radio.	Très élevée
SINR, RSRQ, RSSI bruts	Permet de valider physiquement l'alerte.	Élevée
Historique court	Identifie les épisodes persistants.	Élevée
Statut de dérive	Signale un changement de distribution.	Moyenne à élevée

Dans le notebook, un résumé du monitoring indique 200 échantillons analysés, 52 alertes déclenchées, 14 vrais positifs et 38 faux positifs dans une simulation donnée. Cette sortie montre que le choix du seuil ou de la logique de vote peut fortement influencer le nombre d'alertes. Une calibration opérationnelle est donc nécessaire avant tout déploiement.

4.20 Analyse critique des résultats

4.20.1 Pourquoi les modèles d'arbres dominent

Les modèles d'arbres obtiennent les meilleures performances. Ce résultat est cohérent avec la nature du problème. Les labels sont construits par des conditions à seuils sur des variables tabulaires. Les arbres de décision sont précisément conçus pour apprendre ce type de frontière. Random Forest et XGBoost peuvent donc reconstruire efficacement les règles de décision.

Le 1D-CNN, en revanche, est moins naturellement adapté à ce problème. Il impose une structure locale entre les variables, alors que les KPI n'ont pas nécessairement une relation de voisinage comparable à celle des pixels d'une image ou des échantillons d'un signal temporel. Cette différence explique probablement son F1-score plus faible.

4.20.2 Risque de réapprentissage de la règle d'étiquetage

La limite principale du chapitre est le risque de réapprentissage de la règle d'étiquetage. Comme la cible est construite à partir du SINR et du RSRQ, un modèle recevant ces mêmes variables en entrée peut apprendre à reproduire la règle. La performance devient alors une mesure de capacité à retrouver la règle, non une preuve de détection d'interférences réelles indépendantes.

La Fig. 4.30 résume ce risque et les corrections recommandées. Cette figure est volontairement placée dans la discussion, car elle constitue un point méthodologique majeur du mémoire.

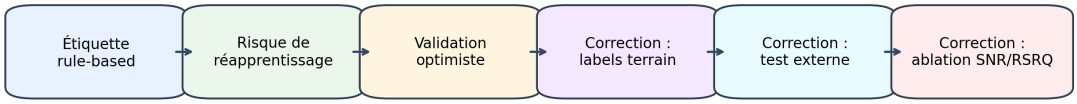


Fig. 4.30 – Risque méthodologique lié à l'étiquetage rule-based et corrections possibles.

Pour renforcer le travail, trois tests complémentaires sont recommandés. Le premier consiste à évaluer le modèle sur un jeu externe annoté indépendamment. Le deuxième consiste à réaliser une étude d'ablation, en retirant temporairement SINR et RSRQ pour voir si les autres KPI conservent un pouvoir prédictif. Le troisième consiste à comparer les prédictions aux événements réels du réseau ou à des mesures spectrales indépendantes.

4.20.3 Interprétation opérationnelle

Malgré cette limite, le pipeline apporte une valeur opérationnelle. Il montre comment transformer des KPI radio en alertes interprétables. Il offre aussi un cadre pour intégrer plusieurs modèles et comparer leurs décisions. Cette architecture pourra être réutilisée lorsque des données réelles annotées seront disponibles.

Le résultat le plus important n'est donc pas uniquement l'accuracy élevée, mais la construction d'une chaîne complète : données, prétraitement, apprentissage, évaluation, explicabilité, vote et monitoring. Cette chaîne est transférable à des datasets réels, à condition de remplacer les étiquettes rule-based par des labels plus fiables.

4.21 Limites identifiées

Les limites du travail peuvent être regroupées en cinq familles. La première concerne la nature synthétique du dataset utilisé dans l'exécution analysée. La deuxième concerne l'étiquetage par règles. La troisième concerne l'absence de validation externe. La quatrième concerne l'absence de modélisation temporelle. La cinquième concerne l'usage du 1D-CNN sur un vecteur de KPI tabulaires.

Tab. 4.19 – Limites principales du chapitre et pistes de correction.

Limite	Conséquence	Correction recommandée
Dataset synthétique	Les résultats ne démontrent pas une généralisation terrain.	Utiliser plusieurs fichiers de mesures réelles et les archiver.
Étiquetage rule-based	Les modèles peuvent réapprendre les règles de seuil.	Obtenir des labels experts ou indépendants.
Absence de validation externe	Le test provient de la même distribution que l'entraînement.	Tester sur une autre session, une autre zone ou un autre opérateur.
Pas de temporalité	Les épisodes persistants ne sont pas modélisés.	Ajouter des fenêtres temporelles et des modèles séquentiels.
1D-CNN sur KPI tabulaires	La convolution dépend d'un ordre artificiel des variables.	Utiliser soit des modèles tabulaires, soit des séquences temporelles réelles.
Seuil de décision fixe	Le compromis précision/rappel peut être sous-optimal.	Optimiser le seuil selon le coût des erreurs.

Limite	Conséquence	Correction recommandée
SMOTE synthétique	Les exemples générés peuvent ne pas refléter les cas terrain.	Comparer avec class weights et données réelles supplémentaires.

Ces limites ne remettent pas en cause la cohérence du pipeline, mais elles définissent clairement les conditions nécessaires pour transformer la preuve de concept en système validé.

4.22 Perspectives d'amélioration

4.22.1 Validation scientifique renforcée

La perspective prioritaire est la validation sur données réelles. Il faudrait collecter des mesures dans plusieurs conditions : zones différentes, heures différentes, charges réseau différentes, et si possible plusieurs opérateurs ou plusieurs cellules. Les événements d'interférence devraient être confirmés par une source indépendante.

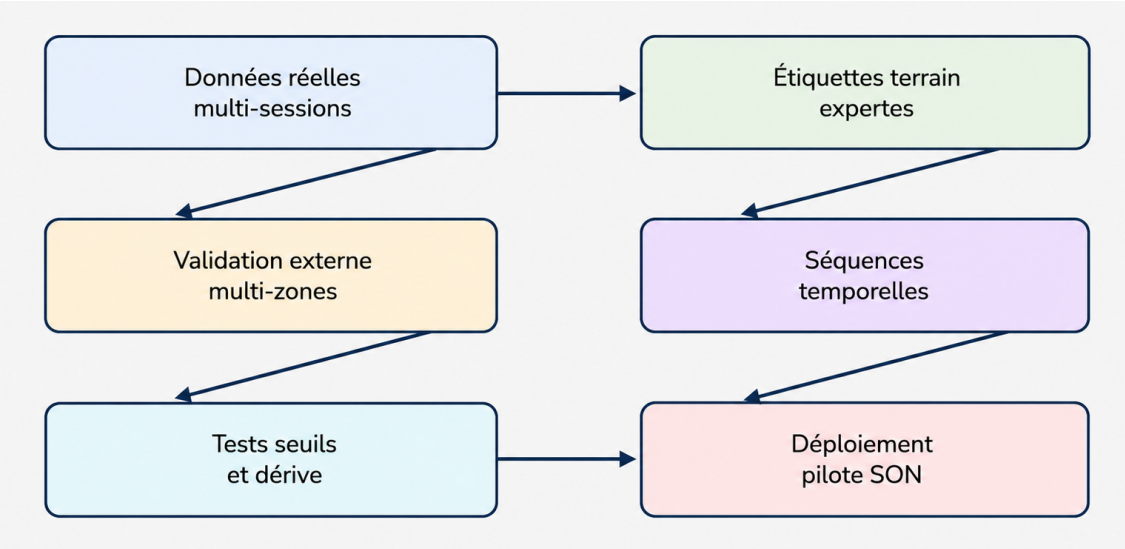


Fig. 4.31 – Feuille de route proposée pour renforcer la validation scientifique.

Une validation robuste devrait inclure un test externe. Par exemple, le modèle serait entraîné sur une session et testé sur une autre session réalisée à une autre période. Cette démarche permettrait de vérifier si le modèle apprend des mécanismes radio généraux ou seulement les particularités du dataset d'origine.

4.22.2 Modèles temporels

Les interférences réelles ont souvent une durée, une répétition ou une dynamique. Le modèle actuel traite chaque observation indépendamment. Une amélioration naturelle consiste à créer des fenêtres temporelles contenant plusieurs mesures successives. Un modèle LSTM, GRU, Temporal CNN ou Transformer pourrait alors apprendre la dynamique de la dégradation.

Tab. 4.20 – Extensions possibles vers des modèles temporels.

Approche	Avantage	Limite
Fenêtre glissante + RF	Simple et compatible avec les modèles actuels.	Résume la temporalité par des statistiques.
LSTM/GRU	Capable d'apprendre une dynamique temporelle.	Besoin de plus de données et d'un réglage plus délicat.
Temporal CNN	Convolutions sur des séquences réelles de KPI.	Choix de la longueur de fenêtre important.
Transformer	Modélisation flexible des dépendances longues.	Plus coûteux et exigeant en données.

4.22.3 Optimisation des seuils et calibration

Une autre perspective consiste à optimiser le seuil de décision. Le seuil par défaut de 0,5 n'est pas forcément optimal. Dans un réseau où les fausses alertes coûtent cher, il peut être préférable d'augmenter le seuil. Dans un réseau critique où les interférences non détectées sont très pénalisantes, il peut être préférable de réduire le seuil.

La calibration des probabilités est également importante. Une probabilité de 90 % doit idéalement correspondre à une fréquence réelle proche de 90 % d'interférences. Des méthodes comme Platt scaling ou isotonic regression pourraient être testées sur un jeu de validation.

4.23 Synthèse scientifique du chapitre

Les résultats montrent que les modèles à base d'arbres sont très performants dans la configuration testée. Random Forest atteint une classification parfaite sur le jeu de test synthétique. XGBoost obtient une performance presque équivalente. Le 1D-CNN est moins précis mais conserve une AUC élevée, ce qui suggère une séparation probabiliste correcte.

L'analyse d'explicabilité confirme le rôle central du SINR, du RSRQ et de la variable dérivée SNR_CQI_Product. Cette cohérence est satisfaisante car ces variables sont physiquement liées à la qualité radio. Cependant, cette même cohérence rappelle la limite principale du protocole : la cible a été construite à partir de ces variables.

La conclusion scientifique doit donc être formulée avec précision. Le chapitre démontre la

faisabilité d'une chaîne de détection d'interférences 5G par machine learning, intégrant prétraitement, rééquilibrage, comparaison de modèles, explicabilité et vote majoritaire. En revanche, il ne démontre pas encore une généralisation terrain indépendante. Cette généralisation nécessitera des mesures réelles, des labels experts et une validation externe.

4.24 Conclusion

Ce chapitre a présenté une version complète et enrichie de l'implémentation du système de détection des interférences 5G. L'ensemble du workflow a été détaillé, depuis l'importation des données jusqu'à l'explication des décisions. Les figures de pipeline ont permis de clarifier la logique de traitement, la construction des labels, le prétraitement, l'entraînement des modèles, l'évaluation et le monitoring opérationnel.

Sur le plan expérimental, Random Forest et XGBoost obtiennent les meilleures performances sur le jeu de test synthétique. Le 1D-CNN fournit des résultats inférieurs en précision, mais reste intéressant pour l'explicabilité par Grad-CAM. Le vote majoritaire constitue une base utile pour une décision plus robuste, surtout dans une perspective de tableau de bord réseau.

Le chapitre met également en évidence une limite méthodologique importante : les labels étant générés par une règle sur les KPI, les modèles peuvent réapprendre cette règle. Cette limite a été explicitement discutée afin d'éviter une interprétation excessive des performances. Les perspectives proposées portent donc sur l'utilisation de données réelles, la validation externe, l'intégration de la dimension temporelle, la calibration des seuils et l'amélioration du protocole d'explicabilité.

En l'état, le travail constitue une preuve de concept solide et pédagogiquement complète. Il fournit une base exploitable pour un mémoire, mais aussi une architecture réutilisable pour des expérimentations futures sur des mesures 5G réelles.

Conclusion générale et perspectives

Ce mémoire a porté sur la détection des interférences dans les réseaux 5G à l'aide de techniques de Machine Learning. Le point de départ du travail est le constat que les environnements radio modernes sont de plus en plus denses, hétérogènes et dynamiques. Dans ces conditions, les interférences peuvent dégrader la qualité de service en réduisant le débit, en augmentant les pertes, en diminuant la qualité radio et en perturbant la stabilité de la connexion. La détection précoce de ces situations constitue donc un enjeu important pour la supervision et l'optimisation des réseaux mobiles.

Le premier chapitre a présenté les éléments nécessaires à la compréhension des réseaux 5G : évolution des générations cellulaires, cas d'usage, bandes de fréquences, architecture radio et coeur, technologies clés et indicateurs de qualité. Il a également mis en évidence le lien entre les phénomènes d'interférences et les KPIs radio tels que SINR, RSRP, RSRQ, RSSI, CQI et BLER. Cette base théorique a permis de justifier le choix d'une approche orientée données.

Le deuxième chapitre a introduit les méthodes de Machine Learning appliquées aux réseaux 5G. Il a montré que les données télécoms peuvent être transformées en variables exploitables par des modèles de classification, de détection d'anomalies ou d'aide à la décision. Les méthodes ensemblistes, le deep learning et les outils d'explicabilité ont été replacés dans une logique de supervision intelligente, cohérente avec les tendances d'automatisation des réseaux, notamment 5G-Advanced, O-RAN et zero-touch management.

Le troisième chapitre a décrit le dataset, les variables utilisées et la méthodologie de préparation des données. L'étiquetage repose sur une lecture physique simplifiée des KPIs radio. Le nettoyage, l'encodage, la normalisation, le découpage train/test, le rééquilibrage par SMOTE et la sélection de variables ont été présentés comme des étapes indispensables pour éviter les biais d'évaluation et améliorer la reproductibilité.

Le quatrième chapitre a présenté l'implémentation expérimentale. Trois modèles ont été étudiés : Random Forest, XGBoost et 1D-CNN. Les performances ont été évaluées à l'aide de métriques complémentaires : accuracy, précision, rappel, F1-score, matrices de confusion et courbes ROC. L'analyse a montré que les KPIs contiennent une information utile pour distinguer les états normaux et les états interférés. L'explicabilité par importance des variables, SHAP et Grad-CAM a permis de relier les décisions des modèles aux indicateurs radio les plus influents.

La contribution principale de ce travail est donc la construction d'une chaîne complète et reproductible de détection des interférences 5G, depuis l'analyse des KPIs jusqu'à l'interprétation des décisions. Le mémoire insiste sur un point méthodologique important : une bonne performance numérique ne suffit pas. Pour être exploitable dans un contexte réseau, un modèle doit aussi être interprétable, cohérent avec la physique radio, documenté et validé sur des données représentatives.

Plusieurs limites doivent cependant être soulignées. La validation expérimentale dépend du dataset disponible et des règles d'étiquetage utilisées. Les résultats doivent donc être interprétés comme une preuve de concept méthodologique plutôt que comme une validation opérationnelle définitive. De plus, l'approche exploite principalement des observations tabulaires et ne modélise pas encore pleinement la dynamique temporelle des KPIs, la mobilité des utilisateurs ou le contexte multi-cellules.

Les perspectives de ce travail sont nombreuses. Premièrement, il serait nécessaire d'enrichir la base de données avec des mesures réelles collectées sur plusieurs zones, bandes de fréquences, cellules et scénarios de trafic. Deuxièmement, l'étiquetage devrait être confronté à des annotations d'experts ou à des scénarios contrôlés d'interférences. Troisièmement, des modèles temporels tels que LSTM, GRU ou Transformers pourraient être étudiés afin de capturer l'évolution des KPIs dans le temps. Quatrièmement, les seuils de décision devraient être calibrés selon les contraintes opérationnelles afin d'équilibrer fausses alertes et détections manquées. Enfin, la chaîne proposée pourrait être intégrée dans un tableau de bord de monitoring ou dans une boucle O-RAN orientée automatisation.

En conclusion, ce mémoire montre que le Machine Learning constitue une approche pertinente pour la détection des interférences dans les réseaux 5G. L'association entre connaissance radio, préparation rigoureuse des données, modèles de classification et outils d'explicabilité ouvre la voie à des systèmes de supervision plus intelligents, plus adaptatifs et plus utiles pour l'exploitation des réseaux mobiles modernes.

Bibliographie

[1] *Téléphonie mobile - 3G et 4G expliquées*. Accessed : Apr. 04, 2026.

[2] HTCN. *Les différents réseaux mobiles : EDGE, 3G, 4G, 5G*. Accessed : Apr. 04, 2026.

[3] Salhane, I., Rifi, M. *La technologie 5G pour les objets connectés*. In : Colloque sur les Objets et systèmes Connectés (COC'2021). 2021. Accessed : Apr. 03, 2026.

[4] *La MFS considère la technologie sans fil 5G comme un agent perturbateur de l'industrie*. 2026. Accessed : Apr. 04, 2026.

[5] Dabbous, M. *Système antennaire en bandes Sub-6 GHz et millimétrique pour petites cellules 5G*. Université Côte d'Azur. 2025. Accessed : Apr. 04, 2026.

[6] School of Computing, Sathyabama Institute of Science, Technology, others. *Design of an Integrated mm-Wave and Sub-6GHz Antenna for 5G Mobile Devices*. Journal of Nano- and Electronic Physics. vol. 15. no. 4. pp. 04027-1–04027-5. 2023.

[7] everything RF. *What is Sub-6 GHz 5G ?*. Accessed : Apr. 04, 2026.

[8] Holma, Harri, Toskala, Antti, Nakamura, Takehiro. *5G Technology : 3GPP New Radio*. Wiley. 2024.

[9] Launay, Frédéric. *NG-RAN and 5G-NR : 5G Radio Access Network and Radio Interface*. ISTE Ltd / John Wiley and Sons Inc. 2021.

[10] Gonzague. *5G NSA, 5G SA, quelle est la différence ?*. Accessed : Apr. 05, 2026.

[11] Moutaly, E. *Étude d'une architecture radio-sur-fibre pour le fronthaul 5G*. 2020.

[12] Grandmetric. *5G Core Network (5GC) Functions*. Accessed : Apr. 04, 2026.

[13] El Bouhani, Abdelbasset. *Étude des scénarios de clustering des réseaux 4G de TT pour la migration vers une architecture 5G Cloud-RAN*. Memoire Online. Accessed : Apr. 04, 2026.

[14] Siddiqui, M. U. A., Qamar, F., Ahmed, F., Nguyen, Q. N., Hassan, R. *Interference Management in 5G and Beyond Network : Requirements, Challenges and Future Directions*. IEEE Access. vol. 9. pp. 68932–68965. 2021.

[15] Reza, D. A., Amaelia, R., Saputra, S. J., Wijayanti, H. A., Lubis, L. K. *Interference Analysis between 5G Network and LoRa Technology in Wireless Communication System*. International Journal. vol. 10. no. 2. 2023.

[16] Trabelsi, N., Fourati, L. C., Chen, C. S. *Interference Management in 5G and Beyond Networks*. 2024.

[17] Afroz, F., Subramanian, R., Heidary, R., Sandrasegaran, K., Ahmed, S. *SINR, RSRP, RSSI and RSRQ Measurements in Long Term Evolution Networks*. International Journal of Wireless & Mobile Networks. vol. 7. no. 4. pp. 113–123. 2015.

[18] 3GPP. *TS 36.214: LTE ; Evolved Universal Terrestrial Radio Access (E-UTRA) ; Physical Layer Measurements*. no. 3GPP TS 36.214 V15.3.0 Release 15. 2018.

[19] Aniru, M. A., Omoavowere, E. J. *Investigation and Analysis of End-User Service Integrity in LTE/5G Fixed Wireless Access Networks in Macrocellular Environments*. 2024.

[20] Wang, J., Rouil, R. A. *BLER Performance Evaluation of LTE Device-to-Device Communications*. no. NIST IR 8157. 2016.

[21] O'Shea, Timothy, Hoydis, Jakob. *The Convergence of Machine Learning and Communications*. ITU Journal : ICT Discoveries. vol. 1. 2018.

[22] Birmingham City University. *A Review of Machine Learning Use-Cases in Telecommunications*. 2023.

[23] *Machine Learning and Wireless Communications*. Cambridge University Press. 2022.

[24] 3GPP. *AI/ML for NR Air Interface*. 2023. 3GPP technology overview, Release 18 AI/ML study and work items.

[25] O-RAN Alliance. *O-RAN Architecture and Resources*. 2026. O-RAN resources describing RIC, xApps, rApps and intelligent RAN automation.

[26] ETSI. *Industry Specification Group on Zero touch network and Service Management*. 2026. ETSI ZSM overview of end-to-end automation for networks and services.

[27] Lin, Xingqin. *An Overview of the 3GPP Study on Artificial Intelligence for 5G New Radio*. arXiv preprint arXiv :2308.05315. 2023.

[28] Lin, Xingqin. *Artificial Intelligence in 3GPP 5G-Advanced : A Survey*. arXiv preprint arXiv :2305.05092. 2023.

[29] 3GPP. *TS 38.215: NR ; Physical Layer Measurements*. no. 3GPP TS 38.215 version 17.4.0 Release 17. 2024.

[30] Raca, Darijo, Leahy, Dylan, Sreenan, Cormac J., Quinlan, Jason J. *Beyond Throughput, the Next Generation : A 5G Dataset with Channel and Context Metrics*. In : Proceedings of the 11th ACM Multimedia Systems Conference. pp. 303–308. 2020.

[31] Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron. *Deep Learning*. MIT Press. 2016.

[32] Jha, A., Kumar, R., others. *Optimizing 5G Sustainable Networks Performance in Urban Areas Using Machine Learning Techniques*. International Journal of Communication Systems. Wiley. 2025.

[33] Varotto, M., Valentin, S., Tomasin, S. *Detecting 5G Signal Jammers Using Spectrograms with Supervised and Unsupervised Learning*. arXiv preprint arXiv :2405.10331. 2024.

[34] Sesia, Stefania, Toufik, Issam, Baker, Matthew. *LTE - The UMTS Long Term Evolution : From Theory to Practice*. John Wiley & Sons. 2011.

[35] Breiman, Leo. *Random Forests*. Machine Learning. vol. 45. no. 1. pp. 5–32. 2001.

[36] Chen, Tianqi, Guestrin, Carlos. *XGBoost : A Scalable Tree Boosting System*. In : Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. 2016.

[37] LeCun, Yann, Bengio, Yoshua, Hinton, Geoffrey. *Deep Learning*. Nature. vol. 521. pp. 436–444. 2015.

[38] Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media. 2019.

[39] Han, Jiawei, Kamber, Micheline, Pei, Jian. *Data Mining : Concepts and Techniques*. Morgan Kaufmann. 2011.

[40] Lundberg, Scott M., Lee, Su-In. *A Unified Approach to Interpreting Model Predictions*. In : Advances in Neural Information Processing Systems 30. pp. 4765–4774. 2017.

[41] Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, Matthieu, Duchesnay, Édouard. *Scikit-learn : Machine Learning in Python*. Journal of Machine Learning Research. vol. 12. pp. 2825–2830. 2011.

[42] Fawcett, Tom. *An Introduction to ROC Analysis*. Pattern Recognition Letters. vol. 27. no. 8. pp. 861–874. 2006.

[43] Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrence O., Kegelmeyer, W. Philip. *SMOTE : Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research. vol. 16. pp. 321–357. 2002.

[44] Lemaître, Guillaume, Nogueira, Fernando, Aridas, Christos K. *Imbalanced-learn : A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*. Journal of Machine Learning Research. vol. 18. no. 17. pp. 1–5. 2017.

[45] Selvaraju, Ramprasaath R., Cogswell, Michael, Das, Abhishek, Vedantam, Ramakrishna, Parikh, Devi, Batra, Dhruv. *Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization*. In : Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626. 2017.

Annexe A

Annexe code : implémentation Python du système de détection

Cette annexe présente les principaux blocs de code utilisés pour construire le pipeline de détection des interférences 5G. Le notebook complet est conservé dans le dossier `sources_codes/` du package LaTeX. Un script Python nettoyé est également fourni dans ce même dossier afin de faciliter la consultation et la réutilisation du travail.

A.1 Import des bibliothèques principales

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.metrics import (classification_report,
    confusion_matrix,
                             accuracy_score, f1_score,
                                roc_auc_score)
from sklearn.ensemble import RandomForestClassifier

from xgboost import XGBClassifier
from imblearn.over_sampling import SMOTE

import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv1D, MaxPooling1D, Dense,
    Dropout, Flatten
```

Listing A.1 – Import des bibliothèques utilisées dans le pipeline 5G.

A.2 Chargement ou génération du dataset

```
import os

file_path = "B_2020.02.27_18.39.27.csv"

if os.path.exists(file_path):
    df = pd.read_csv(file_path)
else:
    n = 1000
    df = pd.DataFrame({
        "SINR": np.random.uniform(-5, 20, n),
        "RSRP": np.random.uniform(-120, -70, n),
        "RSRQ": np.random.uniform(-20, -5, n),
        "RSSI": np.random.uniform(-110, -60, n),
        "DL_bitrate": np.random.uniform(0, 100000, n),
        "UL_bitrate": np.random.uniform(0, 10000, n),
        "CQI": np.random.randint(1, 15, n),
        "Timestamp": pd.date_range(start="2023-01-01", periods=n,
           freq="s")
    })
```

Listing A.2 – Chargement du fichier de mesures ou génération d'un dataset de démonstration.

A.3 Construction d'une étiquette binaire d'interférence

```
def build_interference_label(row):
    weak_quality = (row["SINR"]  < 5) or (row["RSRQ"] < -15)
    low_service = (row["DL_bitrate"] < 15000) or (row["CQI"] < 5)
    high_total_power = row["RSSI"] > -75
    return int((weak_quality and low_service) or (high_total_power
       and row["SINR"]  < 8))

df["interference"] = df.apply(build_interference_label, axis=1)
```

Listing A.3 – Exemple de règle physique simplifiée pour construire une cible binaire.

A.4 Prétraitement et séparation apprentissage-test

```
features = ["SINR", "RSRP", "RSRQ", "RSSI", "DL_bitrate", "
   UL_bitrate", "CQI"]
X = df[features].copy()
y = df["interference"].astype(int)

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.20, random_state=42, stratify=y
)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

smote = SMOTE(random_state=42)
X_train_balanced, y_train_balanced = smote.fit_resample(
   X_train_scaled, y_train)
```

Listing A.4 – Normalisation, séparation des données et rééquilibrage de l'ensemble d'apprentissage.

A.5 Entraînement de Random Forest et XGBoost

```
rf_model = RandomForestClassifier(
    n_estimators=200,
    max_depth=None,
    random_state=42,
    class_weight="balanced"
)
rf_model.fit(X_train_balanced, y_train_balanced)

xgb_model = XGBClassifier(
    n_estimators=200,
    learning_rate=0.05,
    max_depth=4,
    subsample=0.9,
    colsample_bytree=0.9,
    eval_metric="logloss",
    random_state=42
)
```

```
xgb_model.fit(X_train_balanced, y_train_balanced)
```

Listing A.5 – Entraînement de deux modèles à arbres ensemblistes.

A.6 Architecture 1D-CNN

```
X_train_cnn = X_train_balanced.reshape((X_train_balanced.shape[0],
    X_train_balanced.shape[1], 1))
X_test_cnn = X_test_scaled.reshape((X_test_scaled.shape[0],
    X_test_scaled.shape[1], 1))

cnn_model = Sequential([
    Conv1D(filters=32, kernel_size=3, activation="relu",
        input_shape=(X_train_cnn.shape[1], 1)),
    MaxPooling1D(pool_size=2),
    Dropout(0.25),
    Conv1D(filters=64, kernel_size=2, activation="relu"),
    Flatten(),
    Dense(64, activation="relu"),
    Dropout(0.25),
    Dense(1, activation="sigmoid")
])

cnn_model.compile(optimizer="adam", loss="binary_crossentropy",
    metrics=["accuracy"])
cnn_model.fit(X_train_cnn, y_train_balanced, epochs=30, batch_size
    =32, validation_split=0.2)
```

Listing A.6 – Exemple d'architecture 1D-CNN compacte pour données KPI.

A.7 Évaluation et explicabilité

```
y_pred = rf_model.predict(X_test_scaled)
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))

importance = pd.Series(rf_model.feature_importances_, index=
    features)
importance = importance.sort_values(ascending=False)
print(importance)
```

```
8
9  # SHAP et Grad-CAM sont utilisés dans le notebook complet pour
       expliquer
10 # respectivement les modèles à arbres et le modèle 1D-CNN.
```

Listing A.7 – Évaluation et analyse d'importance des variables.

A.8 Remarque sur la reproductibilité

Pour garantir la reproductibilité, les séparations de données et les modèles utilisent une valeur `random_state`. Dans une validation industrielle, il est recommandé d'ajouter une séparation temporelle, une validation croisée par scénario radio et une évaluation sur plusieurs fichiers de mesures indépendants.

Annexe B

Annexe données : dictionnaire des variables 5G

Cette annexe synthétise les principales variables utilisées pour la détection des interférences. Elle complète le chapitre 3 en donnant une lecture rapide des KPI, de leur unité et de leur rôle dans la chaîne de Machine Learning.

Tab. B.1 – Dictionnaire synthétique des variables radio utilisées.

Variable	**Unité**	**Rôle dans la détection**
SINR / SINR	dB	Mesure le rapport entre le signal utile et le bruit ou les interférences ; indicateur directement lié à la qualité radio.
RSRP	dBm	Représente la puissance reçue du signal de référence ; utile pour distinguer un problème de couverture d'un problème d'interférence.
RSRQ	dB	Indique la qualité du signal de référence ; sensible à la charge radio et aux interférences.
RSSI	dBm	Mesure la puissance totale reçue ; peut augmenter en présence de signaux interférents.
CQI	1–15	Résume la qualité du canal et influence le choix de modulation et de codage.
DL_bitrate	Kbps	Traduit l'effet pratique de l'état radio sur le service descendant.
UL_bitrate	Kbps	Complète l'analyse du lien montant et peut révéler des dégradations spécifiques.
BLER	%	Mesure le taux d'erreur de blocs ; augmente lorsque la qualité radio se dégrade.
Timestamp	date/heure	Permet d'étudier la dynamique temporelle des épisodes de dégradation.