Université Mohammed-V de Rabat
École Nationale Supérieure dInformatique et dAnalyse des Systèmes

ENSIAS

Rapport de projet S5

Filière : Data science & IoT (IDSIT)

Detection des metaux lourds dans l'eau agricole

Réalisé par :

M. AREKTOUT Mossab

Encadré par :

Pr. MOuMANE Karima

Soutenu le 22 janvier 2026, Devant le jury composé de :

Pr. NAFIL Khalid
Pr. MOUMANE Karima

Année universitaire 2025/2026

Remerciements

Avant tout développement sur ce travail, il m'apparaît opportun de commencer par exprimer mes sincères remerciements à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce projet de fin de semestre.

Je tiens à remercier mon encadrante, Madame **Karima Moumane**, pour avoir supervisé ce projet et pour le cadre de travail qu'elle a mis à ma disposition.

Mes remerciements les plus chaleureux s'adressent à **Amin Bellali**, doctorant, pour son implication remarquable, son soutien technique constant et son partage généreux de connaissances. Sa disponibilité, sa patience et son expertise ont été déterminantes dans l'aboutissement de ce travail. Cette collaboration enrichissante m'a permis d'acquérir de nouvelles compétences et de surmonter les nombreux défis rencontrés.

Enfin, je tiens à exprimer ma profonde reconnaissance envers ma famille pour leur soutien inconditionnel, leurs encouragements constants et leur confiance indéfectible. Leur présence à mes côtés a été une source de motivation tout au long de mon parcours académique.

À tous, je vous adresse mes plus sincères remerciements.

Résumé

La contamination des eaux agricoles par les métaux lourds constitue une menace croissante pour la santé publique et l'environnement. Ce projet présente le développement d'un système intelligent de détection du Chrome (Cr) et de l'Arsenic (As) dans l'eau agricole, basé sur la spectroscopie Raman exaltée de surface (SERS) couplée aux techniques d'apprentissage automatique. L'approche adoptée repose sur une architecture hiérarchique à deux niveaux : un premier modèle classifie le type de contamination (Control/Cr/As), puis des modèles spécialisés quantifient la concentration de chaque métal. Le pipeline de traitement intègre une correction de ligne de base ALS, un lissage Savitzky-Golay, une réduction dimensionnelle par PCA, et une classification par SVM et Random Forest. Les résultats expérimentaux démontrent l'efficacité du système pour la détection et la quantification précise des métaux lourds, offrant une alternative prometteuse aux méthodes analytiques traditionnelles.

Mots-clés : Métaux lourds, SERS, Machine Learning, Classification hiérarchique, Chrome, Arsenic, Qualité de l'eau.

Abstract

Heavy metal contamination in agricultural water poses a growing threat to public health and the environment. This project presents the development of an intelligent system for detecting Chromium (Cr) and Arsenic (As) in agricultural water, based on Surface-Enhanced Raman Spectroscopy (SERS) coupled with machine learning techniques. The adopted approach relies on a two-level hierarchical architecture : a first model classifies the contamination type (Control/Cr/As), then specialized models quantify the concentration of each metal. The processing pipeline includes ALS baseline correction, Savitzky-Golay smoothing, PCA dimensionality reduction, and classification using SVM and Random Forest. Experimental results demonstrate the system's effectiveness for accurate heavy metal detection and quantification, offering a promising alternative to traditional analytical methods.

Keywords : Heavy metals, SERS, Machine Learning, Hierarchical classification, Chromium, Arsenic, Water quality.

Table des matières

Remerciements **3**

Résumé **4**

Abstract **5**

Table des matières **8**

Table des figures **9**

Liste des tableaux **10**

Liste des acronymes **11**

Introduction Générale **12**

1 Étude Bibliographique **13**
1.1 Les Métaux Lourds 13
1.1.1 Définition et caractéristiques 13
1.1.2 Principaux métaux lourds 14
1.1.3 Sources de contamination 14
1.1.4 Impacts sur la santé et l'environnement 15
1.2 La Qualité de l'Eau Agricole 16
1.2.1 Importance de l'eau dans l'agriculture 16
1.2.2 Normes et seuils réglementaires 16
1.2.3 Méthodes traditionnelles de détection 17
1.3 L'Intelligence Artificielle pour la Détection 18
1.3.1 Approches basées sur le Machine Learning 19
1.3.2 Approches basées sur le Deep Learning 19
1.3.3 Synthèse comparative 20
1.3.4 Limites identifiées et positionnement 20
1.3.5 Positionnement de notre travail 21

2 Méthodologie et Outils **22**
2.1 Méthodologie Adoptée 22
2.1.1 Démarche globale du projet 22
2.1.2 Pipeline de traitement des données 24
2.2 Description des Données 25
2.2.1 Source des données 25
2.2.2 Structure des fichiers 25
2.2.3 Classes de concentration 25
2.2.4 Résumé du dataset 28
2.3 Prétraitement des Données 29

2.3.1 Correction de la ligne de base (ALS) . . . 29
2.3.1.0.1 Principe : . . . 29
2.3.2 Lissage spectral (Savitzky–Golay) . . . 30
2.3.2.0.1 Principe : . . . 30
2.3.3 Normalisation (StandardScaler) . . . 30
2.3.4 Réduction de dimensionnalité (PCA) . . . 31
2.3.4.0.1 Avantages de la PCA : . . . 31
2.4 Division des Données et Équilibrage . . . 32
2.4.1 Division Train/Test . . . 32
2.4.2 Équilibrage par SMOTE . . . 32
2.4.2.0.1 Principe de SMOTE : . . . 32
2.4.2.0.2 Avantages : . . . 32
2.4.2.0.3 Remarque importante : . . . 33
2.5 Outils et Technologies Utilisés . . . 33
2.5.1 Langage de programmation . . . 33
2.5.2 Bibliothèques utilisées . . . 34
2.5.3 Environnement de développement . . . 34

3 Conception et Modélisation 36
3.1 Architecture du Système de Détection . . . 36
3.1.1 Approche hiérarchique . . . 36
3.1.2 Flux de prédiction . . . 37
3.1.2.0.1 Étape 1 : Prétraitement du spectre . . . 38
3.1.2.0.2 Étape 2 : Classification du type de métal (Modèle 1) . . . 38
3.1.2.0.3 Étape 3 : Classification de la concentration (Modèle 2a ou 2b) . . . 38
3.1.2.0.4 Étape 4 : Résultat final . . . 38
3.1.3 Avantages de larchitecture proposée . . . 39
3.2 Conception des Modèles . . . 39
3.2.1 Modèle 1 : Classification du type de métal . . . 39
3.2.2 Modèle 2a : Classification de la concentration (Chrome) . . . 40
3.2.3 Modèle 2b : Classification de la concentration (Arsenic) . . . 40
3.2.3.1 Résumé comparatif des modèles . . . 41
3.3 Algorithmes de Classification . . . 41
3.3.1 Support Vector Machine (SVM) . . . 41
3.3.2 Random Forest . . . 42
3.3.3 K-Nearest Neighbors (KNN) . . . 43
3.3.4 Gradient Boosting . . . 43
3.3.5 Justification des choix . . . 44
3.4 Métriques d'Évaluation . . . 45
3.4.1 Accuracy (Exactitude) . . . 45
3.4.2 Precision et Recall . . . 45
3.4.3 F1-Score . . . 46
3.4.4 Matrice de confusion . . . 46
3.4.5 Validation croisée . . . 47
3.4.6 Résumé des métriques utilisées . . . 47

4 Résultats et Discussion 49
4.1 Résultats du Modèle 1 : Classification du Type de Métal . . . 49
4.1.1 Comparaison des algorithmes . . . 49
4.1.2 Matrice de confusion . . . 50
4.1.3 Métriques de performance détaillées . . . 50

4.2 Résultats du Modèle 2a : Concentration du Chrome 51
4.2.1 Performance de classification 51
4.2.1.1 Analyse des erreurs de prédiction 51
4.2.2 Matrice de confusion 51
4.3 Résultats du Modèle 2b : Classification de la Concentration de lArsenic 52
4.3.1 Performance de classification 52
4.3.1.1 Analyse des erreurs de prédiction 53
4.3.1.2 Matrice de confusion 53
4.4 Évaluation du Système Hiérarchique Complet 54
4.4.1 Test du pipeline de bout en bout 54
4.4.2 Robustesse du système 55
4.5 Discussion 55
4.5.1 Interprétation des résultats 55
4.5.2 Limites et perspectives damélioration 56

Conclusion Générale **58**

Bibliographie **59**

Table des figures

1.1 Position des principaux métaux lourds dans le tableau périodique des éléments . 14

2.1 Méthodologie globale 23
2.2 Pipeline de traitement des données 24
2.3 Chrome hexavalent (Cr^{6+}) : Control, 68 pM, 68 nM et 68 ţM. 28
2.4 Arsenic (As) : Control, 0.05 nM, 50 nM et 50 ţM. 28
2.5 Exemples de spectres SERS pour différentes concentrations de Chrome et Arsenic 28
2.6 Effet du prétraitement sur un spectre SERS (a) Avant : spectre brut avec ligne de base (b) Après : spectre corrigé et lissé 30
2.7 Analyse de la variance expliquée par PCA 31
2.8 SMOTE 33

3.1 Approche hiérarchique 37
3.2 Flux de prédiction détaillé 38
3.3 Visualisation t-SNE des classes 40
3.4 Exemple de matrice de confusion 47

4.1 Matrice de confusion du Modèle 1 (Classification du type de métal) 50
4.2 Métriques par classe pour le Modèle 1 50
4.3 Matrice de confusion du Modèle 2a (Concentration du Chrome) 52
4.4 Matrice de confusion du Modèle 2b (Concentration de lArsenic) 54
4.5 Performance des modèles 55

Liste des tableaux

1.1 Principaux métaux lourds, leurs sources et leur toxicité 14
1.2 Concentrations maximales admissibles des métaux lourds dans les eaux d'irrigation selon l'OMS 17
1.3 Synthèse comparative des méthodes de détection des métaux lourds 20

2.1 Description du dataset utilisé 25
2.2 Structure des colonnes des fichiers de spectres SERS 25
2.3 Classes de concentration pour le Chrome (Cr) 26
2.4 Classes de concentration pour l'Arsenic (As) 27
2.5 Résumé global du dataset 28
2.6 Paramètres de la correction ALS 29
2.7 Paramètres du filtre Savitzky–Golay 30
2.8 Paramètres de la PCA 31
2.9 Répartition des ensembles d'apprentissage et de test 32
2.10 Avantages du langage Python 33
2.11 Environnement de développement 34

3.1 Comparaison des trois modèles 41
3.2 Comparaison des algorithmes de classification 44

4.1 Comparaison des algorithmes pour le Modèle 1 49
4.2 Résultats du Modèle 2a (Classification de la concentration du Chrome) 51
4.3 Résultats du Modèle 2b (Classification de la concentration de lArsenic) 53

Liste des acronymes

ALS	*Asymmetric Least Squares (Moindres Carrés Asymétriques)*
As	*Arsenic*
Cd	*Cadmium*
Cr	*Chrome (Chromium)*
CNN	*Convolutional Neural Network*
DL	*Deep Learning (Apprentissage Profond)*
Hg	*Mercure (Mercury)*
IA	*Intelligence Artificielle*
KNN	*K-Nearest Neighbors*
ML	*Machine Learning*
OMS	*Organisation Mondiale de la Santé*
PCA	*Principal Component Analysis*
RF	*Random Forest*
SERS	*Surface-Enhanced Raman Spectroscopy*
SMOTE	*Synthetic Minority Over-sampling Technique*
SVM	*Support Vector Machine*
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
WHO	*World Health Organization*

Introduction Générale

L'agriculture constitue l'un des piliers fondamentaux de l'économie mondiale et de la sécurité alimentaire. Cependant, la qualité de l'eau utilisée pour l'irrigation représente un enjeu majeur pour la santé publique et l'environnement. Parmi les contaminants les plus préoccupants, les métaux lourds (plomb, mercure, etc.) occupent une place centrale en raison de leur toxicité, leur persistance dans l'environnement et leur capacité à s'accumuler dans la chaîne alimentaire.

La contamination des eaux agricoles par les métaux lourds provient de multiples sources : activités industrielles, exploitation minière, utilisation excessive d'engrais et de pesticides, ou encore rejets urbains. Ces polluants, une fois présents dans l'eau d'irrigation, peuvent être absorbés par les cultures, compromettant ainsi la qualité des produits agricoles et posant des risques sanitaires pour les consommateurs.

Face à cette problématique, les méthodes traditionnelles d'analyse des métaux lourds (spectrométrie d'absorption atomique, ICP-MS) présentent certaines limitations : coûts élevés, nécessité d'équipements sophistiqués, délais d'analyse importants et besoin de personnel qualifié. Ces contraintes rendent difficile la mise en place d'un suivi régulier et accessible, particulièrement dans les régions à ressources limitées.

C'est dans ce contexte que l'intelligence artificielle (IA) émerge comme une solution prometteuse. Les techniques d'apprentissage automatique et d'apprentissage profond offrent la possibilité de développer des systèmes de détection rapides, précis et économiques. En exploitant des données issues de capteurs ou d'analyses spectrales, ces modèles peuvent prédire la présence et la concentration de métaux lourds avec une efficacité remarquable.

Le présent projet s'inscrit dans cette dynamique d'innovation en proposant un système basé sur l'IA pour la détection des métaux lourds dans les eaux agricoles. L'objectif est de concevoir un outil capable d'assister les agriculteurs, les organismes de contrôle et les décideurs dans la surveillance de la qualité de l'eau, contribuant ainsi à une agriculture plus sûre et plus durable.

Chapitre 1

Étude Bibliographique

Introduction

Ce premier chapitre constitue le socle théorique de notre projet. Il vise à présenter les concepts fondamentaux nécessaires à la compréhension de notre travail. Nous aborderons dans un premier temps la problématique des métaux lourds et leurs impacts sur l'environnement et la santé. Ensuite, nous examinerons l'importance de la qualité de l'eau dans le contexte agricole ainsi que les méthodes conventionnelles de détection. Enfin, nous introduirons les concepts de l'intelligence artificielle et son application dans le domaine de la détection des polluants, tout en présentant un état de l'art des travaux connexes.

1.1 Les Métaux Lourds

1.1.1 Définition et caractéristiques

Les métaux lourds, également appelés éléments traces métalliques (ETM), désignent un ensemble d'éléments chimiques caractérisés par une masse volumique élevée, généralement supérieure à 5 g/cmş. Cette appellation regroupe une cinquantaine d'éléments du tableau périodique, bien que seule une vingtaine présente un intérêt toxicologique significatif.

Ces éléments se distinguent par plusieurs caractéristiques communes :

— **Persistance environnementale** : Contrairement aux polluants organiques, les métaux lourds ne se dégradent pas et persistent indéfiniment dans l'environnement.

— **Bioaccumulation** : Ils s'accumulent dans les organismes vivants au fil du temps, avec des concentrations qui augmentent à chaque niveau de la chaîne alimentaire (bioamplification).

— **Toxicité** : Même à faibles concentrations, certains métaux lourds présentent une toxicité importante pour les êtres vivants.

— **Ubiquité** : Ils sont présents naturellement dans l'environnement, mais les activités humaines ont considérablement augmenté leurs concentrations.

1.1.2 Principaux métaux lourds

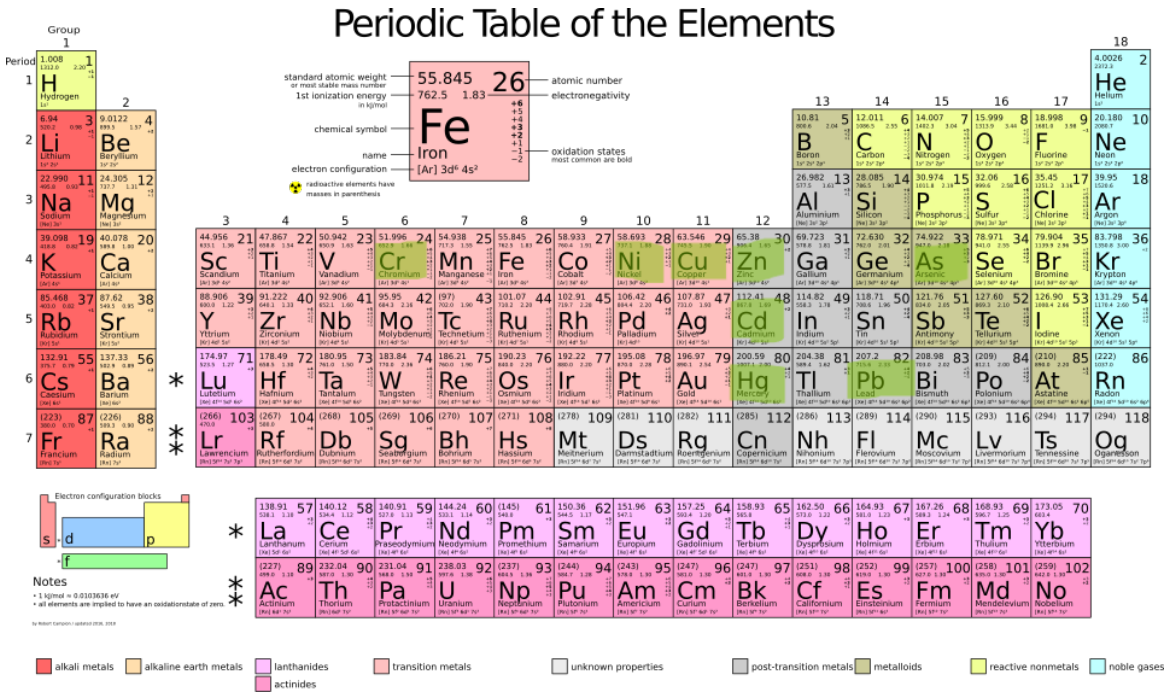

FIGURE 1.1 – Position des principaux métaux lourds dans le tableau périodique des éléments

Métal	Symbole	Densité (g/cm^3)	Sources principales	Toxicité
Plomb	Pb	11,35	Batteries, peintures, canalisations	Neurotoxique, néphrotoxique
Mercure	Hg	13,55	Exploitation minière, combustion	Neurotoxique puissant
Cadmium	Cd	8,65	Engrais phosphatés, industries	Cancérigène, néphrotoxique
Arsenic	As	5,73	Pesticides, exploitation minière	Cancérigène
Chrome	Cr	7,19	Tanneries, galvanoplastie	Cancérigène (Cr VI)
Nickel	Ni	8,90	Industries métallurgiques	Allergène, cancérigène
Cuivre	Cu	8,96	Fongicides, industries	Hépatotoxique à forte dose
Zinc	Zn	7,14	Galvanisation, engrais	Troubles digestifs

TABLE 1.1 – Principaux métaux lourds, leurs sources et leur toxicité

1.1.3 Sources de contamination

La contamination des eaux par les métaux lourds provient de deux catégories de sources :

Sources naturelles

— Érosion des roches et des sols

— Activité volcanique

— Feux de forêt

— Émissions biogéniques

Sources anthropiques

— **Activités industrielles** : métallurgie, sidérurgie, industries chimiques, tanneries, papeteries

— **Activités minières** : extraction et traitement des minerais, drainage minier acide

— **Agriculture** : utilisation d'engrais phosphatés (contenant du cadmium), pesticides, boues d'épuration

— **Transports** : émissions véhiculaires, usure des pneus et des freins

— **Déchets urbains** : incinération des déchets, lixiviats de décharges

— **Rejets domestiques** : eaux usées, produits ménagers

1.1.4 Impacts sur la santé et l'environnement

Impacts sur la santé humaine

L'exposition aux métaux lourds peut se faire par plusieurs voies : ingestion (eau et aliments contaminés), inhalation et contact cutané. Les effets sur la santé varient selon le métal, la dose et la durée d'exposition :

— **Plomb** : saturnisme, troubles neurologiques, retards de développement chez l'enfant, hypertension artérielle, insuffisance rénale

— **Mercure** : maladie de Minamata, troubles neurologiques, malformations congénitales, troubles rénaux

— **Cadmium** : maladie d'Itai-Itai, insuffisance rénale, ostéoporose, cancers pulmonaires

— **Arsenic** : cancers (peau, poumons, vessie), lésions cutanées, troubles cardiovasculaires

— **Chrome hexavalent** : cancers pulmonaires, ulcérations cutanées, troubles respiratoires

Impacts environnementaux

Les métaux lourds affectent l'ensemble des compartiments environnementaux :

— **Sol** : modification de la structure et de la fertilité, perturbation de l'activité microbienne

— **Eau** : contamination des nappes phréatiques et des cours d'eau, perturbation des écosystèmes aquatiques

— **Faune et flore** : bioaccumulation, perturbation des fonctions physiologiques, mortalité

— **Chaîne alimentaire** : biomagnification, contamination des produits agricoles

1.2 La Qualité de l'Eau Agricole

1.2.1 Importance de l'eau dans l'agriculture

L'eau constitue une ressource indispensable pour l'agriculture, représentant environ 70% des prélèvements d'eau douce à l'échelle mondiale. Elle intervient à plusieurs niveaux :

— **Irrigation** : apport d'eau aux cultures pour compenser le déficit pluviométrique

— **Abreuvement du bétail** : besoins en eau des animaux d'élevage

— **Transformation agroalimentaire** : lavage, nettoyage, processus de fabrication

— **Fertilisation** : véhicule des nutriments vers les racines des plantes

La qualité de l'eau d'irrigation influence directement :

— la santé des cultures et leur rendement

— la qualité des produits agricoles

— la santé des sols à long terme

— la sécurité sanitaire des aliments

1.2.2 Normes et seuils réglementaires

Plusieurs organismes internationaux ont établi des normes concernant les concentrations maximales admissibles en métaux lourds dans les eaux d'irrigation.

Normes de l'Organisation Mondiale de la Santé (OMS)

Métal	Concentration maximale (mg/L)
Plomb (Pb)	0,01
Mercure (Hg)	0,001
Cadmium (Cd)	0,003
Arsenic (As)	0,01
Chrome (Cr)	0,05
Nickel (Ni)	0,02
Cuivre (Cu)	2,0
Zinc (Zn)	3,0

TABLE 1.2 – Concentrations maximales admissibles des métaux lourds dans les eaux d'irrigation selon l'OMS

Directives de la FAO pour l'irrigation

La FAO (Organisation des Nations Unies pour l'alimentation et l'agriculture) propose des valeurs guides tenant compte de l'accumulation à long terme des métaux lourds dans les sols ainsi que de leurs effets potentiels sur les cultures et la qualité des produits agricoles.

1.2.3 Méthodes traditionnelles de détection

Plusieurs techniques analytiques sont utilisées pour la détection et la quantification des métaux lourds dans l'eau :

Spectrométrie d'absorption atomique (SAA)

Principe : Mesure de l'absorption de la lumière par les atomes à l'état fondamental.

Avantages : Sensibilité élevée, spécificité.

Limites : Analyse mono-élémentaire, interférences spectrales.

Spectrométrie d'émission optique à plasma à couplage inductif (ICP-OES)

Principe : Excitation des atomes dans un plasma d'argon et mesure de l'émission lumineuse.

Avantages : Analyse multi-élémentaire, large gamme de concentrations.

Limites : Coût d'acquisition et de fonctionnement élevé.

Spectrométrie de masse à plasma à couplage inductif (ICP-MS)

Principe : Ionisation dans un plasma et séparation des ions selon leur rapport masse/charge.

Avantages : Sensibilité exceptionnelle, analyse multi-élémentaire.

Limites : Coût très élevé, maintenance complexe.

Voltampérométrie

Principe : Mesure du courant en fonction du potentiel appliqué.

Avantages : Portabilité, coût modéré.

Limites : Sensibilité aux interférences, préparation d'échantillon.

Limites des méthodes traditionnelles

Limitation	Description
Coût	Équipements onéreux (10 000 à 500 000)
Infrastructure	Nécessité de laboratoires équipés
Personnel	Techniciens qualifiés requis
Temps	Délais d'analyse importants
Accessibilité	Difficile dans les zones rurales/isolées

1.3 L'Intelligence Artificielle pour la Détection

L'application de l'intelligence artificielle dans le domaine de la surveillance environnementale a connu un essor considérable ces dernières années. La détection des métaux lourds dans l'eau, traditionnellement réalisée par des méthodes analytiques coûteuses et chronophages, bénéficie désormais des avancées en apprentissage automatique et apprentissage profond. Cette section présente un état de l'art des travaux de recherche explorant l'utilisation de l'IA pour la détection et la prédiction des concentrations en métaux lourds.

1.3.1 Approches basées sur le Machine Learning

Utilisation des algorithmes d'ensemble

Chen et al. (2020) ont développé un système de prédiction de la qualité de l'eau utilisant les algorithmes Random Forest et Gradient Boosting. En exploitant des paramètres physico-chimiques (pH, conductivité, turbidité, température), leur modèle a atteint une précision de 94% dans la classification de la qualité de l'eau et l'identification des échantillons contaminés.

Ahmed et al. (2021) ont comparé plusieurs algorithmes de Machine Learning pour la prédiction des concentrations en plomb et en arsenic dans les eaux souterraines du Bangladesh. Leurs résultats ont démontré la supériorité du XGBoost avec un coefficient de détermination Rš de 0,91 et une erreur quadratique moyenne (RMSE) de 0,023 mg/L.

Support Vector Machines (SVM)

Li et al. (2019) ont appliqué les SVM pour la classification des niveaux de contamination en cadmium dans les eaux d'irrigation en Chine. En combinant des données spectrales et des paramètres de qualité de l'eau, leur modèle a obtenu une précision de 89,7% avec une sensibilité de 92% pour la détection des échantillons dépassant les seuils réglementaires.

Rashid et al. (2022) ont proposé un modèle SVM optimisé par algorithme génétique pour la détection multi-métaux (Pb, Cd, Cr, Hg). L'optimisation des hyperparamètres a permis d'améliorer la précision de 85% à 93,2%.

1.3.2 Approches basées sur le Deep Learning

Réseaux de neurones profonds (DNN)

Zhang et al. (2021) ont proposé un modèle basé sur les réseaux de neurones profonds pour la détection du plomb et du cadmium dans l'eau potable. Leur architecture, composée de six couches cachées avec régularisation dropout, a atteint une précision de 96,5% avec un temps de prédiction inférieur à une seconde.

Mohammad et al. (2022) ont développé un DNN pour la prédiction simultanée de huit métaux lourds. L'utilisation de la normalisation par lots (batch normalization) et de l'optimiseur Adam a permis d'obtenir une erreur moyenne absolue (MAE) de 0,008 mg/L.

Réseaux de neurones convolutifs (CNN)

Kumar et al. (2022) ont combiné la spectroscopie UV-visible avec des CNN pour la détection multi-métaux. Les spectres d'absorption ont été traités comme des images unidimensionnelles, permettant au modèle d'extraire automatiquement les caractéristiques pertinentes. Cette approche a permis d'identifier simultanément six métaux lourds avec une précision moyenne de 92%.

1.3.3 Synthèse comparative

Auteurs	**Année**	**Méthode**	**Métaux ciblés**	**Performance**
Chen et al.	2020	RF, Gradient Boosting	Multiple	94% accuracy
Li et al.	2019	SVM	Cd	89,7% accuracy
Ahmed et al.	2021	XGBoost	Pb, As	$R^2 = 0{,}91$
Zhang et al.	2021	DNN	Pb, Cd	96,5% accuracy
Kumar et al.	2022	CNN + Spectroscopie	6 métaux	92% accuracy
Wang et al.	2021	LSTM	Multiple	87% accuracy
Liu et al.	2022	CNN-LSTM	Multiple	95,3% accuracy
Zhao et al.	2023	Attention NN	Multiple	94,8% accuracy
Wang et al.	2023	IoT + ML	Multiple	Sensibilité 0,001 mg/L

TABLE 1.3 – Synthèse comparative des méthodes de détection des métaux lourds

1.3.4 Limites identifiées et positionnement

Malgré les avancées significatives réalisées dans la détection des métaux lourds, plusieurs limites persistent dans les travaux existants :

— **Généralisation** : Les modèles sont souvent entraînés sur des données locales, ce qui limite leur capacité de généralisation à d'autres contextes géographiques ou environnementaux.

— **Disponibilité des données** : Il existe un manque de jeux de données publics et standardisés pour l'évaluation et la comparaison des différentes approches.

— **Interprétabilité** : Les modèles basés sur le Deep Learning sont souvent considérés comme des ń boîtes noires ż, rendant difficile l'explication de leurs décisions.

— **Déploiement** : Peu d'études abordent l'implémentation pratique de ces modèles en conditions réelles, notamment en milieu agricole.

1.3.5 Positionnement de notre travail

Notre projet s'inscrit dans la continuité de ces travaux en proposant :

— une approche comparative de plusieurs modèles de *Machine Learning* et de *Deep Learning* ;

— un focus particulier sur l'analyse des eaux d'irrigation agricole ;

— une orientation vers le développement d'une solution accessible, fiable et facilement déployable en environnement réel.

Conclusion

Ce chapitre a permis de poser les bases théoriques nécessaires à la compréhension de notre projet. Nous avons exploré la problématique des métaux lourds, leurs sources, leurs impacts et les méthodes traditionnelles de détection. L'état de l'art des travaux utilisant l'intelligence artificielle démontre le potentiel de ces approches pour développer des solutions de détection innovantes, tout en révélant certaines limites que notre travail cherche à adresser.

Le chapitre suivant présentera la méthodologie adoptée ainsi que les outils et technologies utilisés pour la réalisation de notre système de détection.

Chapitre 2

Méthodologie et Outils

Introduction

Après avoir présenté dans le chapitre précédent les fondements théoriques relatifs aux métaux lourds, à la qualité de l'eau agricole et à l'état de l'art sur l'application de l'intelligence artificielle dans ce domaine, ce deuxième chapitre expose la méthodologie adoptée pour la réalisation de notre système. Nous détaillerons la démarche globale suivie, les données utilisées, les étapes de prétraitement ainsi que les outils et technologies employés pour mener à bien ce projet.

2.1 Méthodologie Adoptée

2.1.1 Démarche globale du projet

La réalisation de ce projet suit une méthodologie structurée en plusieurs phases successives, conformément aux standards des projets de *Machine Learning* et de *Deep Learning*.

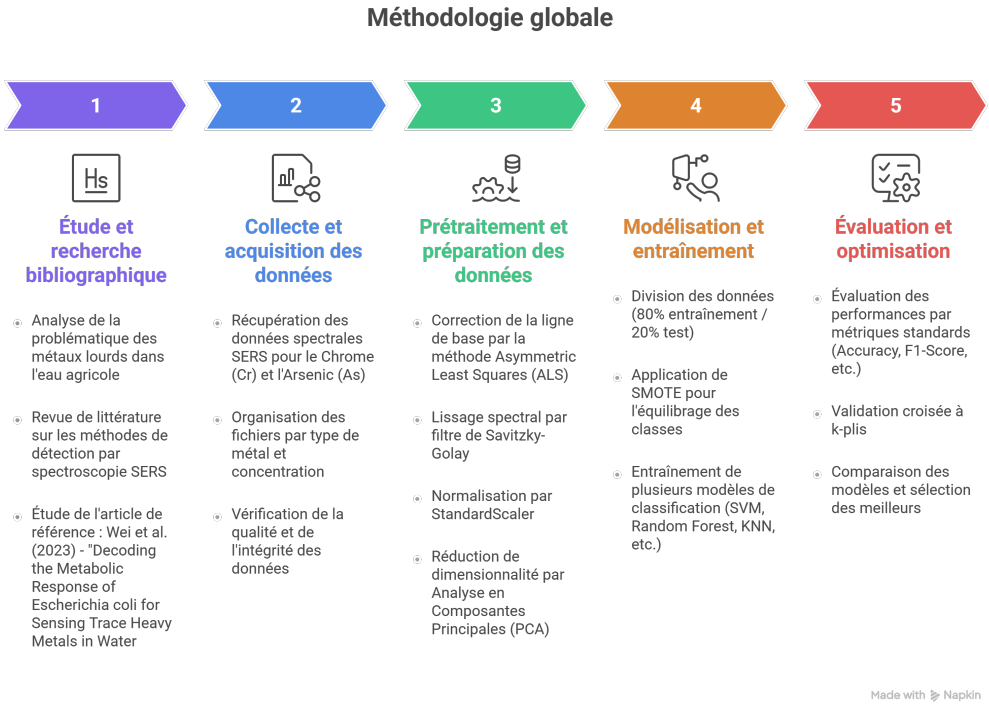

FIGURE 2.1 – Méthodologie globale

Phase 1 : Étude et recherche bibliographique

— Analyse de la problématique des métaux lourds dans l'eau agricole

— Revue de littérature sur les méthodes de détection par spectroscopie SERS

— Étude de l'article de référence : Wei et al. (2023) - "Decoding the Metabolic Response of Escherichia coli for Sensing Trace Heavy Metals in Water"

Phase 2 : Collecte et acquisition des données

— Identification des sources de données disponibles ;

— Sélection du jeu de données approprié ;

— Vérification de la qualité et de la pertinence des données.

Phase 3 : Prétraitement et préparation des données

— Nettoyage des données ;

— Gestion des valeurs manquantes ;

— Normalisation et transformation des variables ;

— Division des données en ensembles d'entraînement, de test et de validation.

Phase 4 : Modélisation et entraînement

— Division des données (80% entraînement / 20% test)

— Application de SMOTE pour l'équilibrage des classes

— Entraînement de plusieurs modèles de classification (SVM, Random Forest, KNN, etc.)

Phase 5 : Évaluation et optimisation

— Évaluation des performances par métriques standards (Accuracy, F1-Score, etc.)

— Validation croisée à k-plis

— Comparaison des modèles et sélection des meilleurs

Phase 6 : Rédaction du rapport et documentation

— Documentation de la solution développée.

— Mise en ligne du projet sur GitHub.

2.1.2 Pipeline de traitement des données

Le pipeline de traitement des données suit un flux séquentiel structuré :

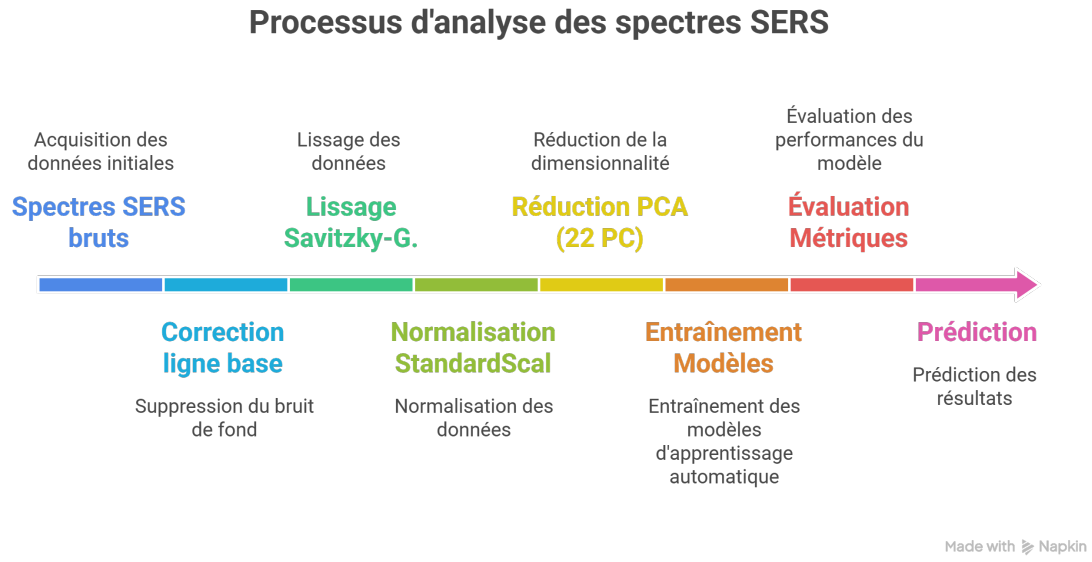

FIGURE 2.2 – Pipeline de traitement des données

Ce pipeline assure une transformation cohérente des données brutes vers un format optimisé pour l'apprentissage automatique, tout en préservant les caractéristiques spectrales discriminantes.

2.2 Description des Données

2.2.1 Source des données

Les données utilisées dans ce projet proviennent de l'étude de Wei et al. (2023) intitulée "Decoding the Metabolic Response of Escherichia coli for Sensing Trace Heavy Metals in Water". Cette étude utilise la spectroscopie SERS (Surface-Enhanced Raman Spectroscopy) couplée à la réponse métabolique de la bactérie Escherichia coli pour détecter les métaux lourds traces dans l'eau.

Caractéristique	**Description**
Source	Wei et al. (2023) Article scientifique
Technique	Spectroscopie SERS (Surface-Enhanced Raman Spectroscopy)
Organisme biosenseur	*Escherichia coli*
Métal ciblé	Chrome hexavalent (Cr^{6+}) et Arsenic (As)
Format des fichiers	Fichiers texte (.txt)

TABLE 2.1 – Description du dataset utilisé

2.2.2 Structure des fichiers

Chaque fichier de données contient des spectres SERS organisés selon la structure suivante :

Colonne	**Index**	**Description**
Colonne 1	0	Coordonnée X (position spatiale)
Colonne 2	1	Coordonnée Y (position spatiale)
Colonne 3	2	Nombre d'onde (Wavenumber) en cm^{-1}
Colonne 4	3	Intensité spectrale (A.U.)

TABLE 2.2 – Structure des colonnes des fichiers de spectres SERS

Chaque spectre comprend **1011 points spectraux** couvrant une plage de nombres d'onde allant de **508.88 à 1640.65 cm^{-1}**.

2.2.3 Classes de concentration

Notre dataset comprend deux métaux lourds avec plusieurs niveaux de concentration.

Classe	**Fichier**	**Concentration**	**Nombre de spectres**
0	00_Control.txt	Control (témoin)	8 000
1	01_0.68_pM.txt	0.68 pM	1 200
2	02_6.8_pM.txt	6.8 pM	1 200
3	03_68_pM.txt	68 pM	1 200
4	04_0.68_nM.txt	0.68 nM	1 200
5	05_6.8_nM.txt	6.8 nM	1 200
6	06_68_nM.txt	68 nM	1 200
7	07_0.68_uM.txt	0.68 μM	1 200
8	08_6.8_uM.txt	6.8 μM	1 200
9	09_68_uM.txt	68 μM	1 200

TABLE 2.3 – Classes de concentration pour le Chrome (Cr)

Classe	**Fichier**	**Concentration**	**Nombre de spectres**
0	00_Control.txt	Control (témoin)	8 000
1	01_5x10^-3_pM.txt	5×10^{-3} pM	1 200
2	02_5x10^-2_pM.txt	5×10^{-2} pM	1 200
3	03_5x10^-4_nM.txt	5×10^{-4} nM	1 200
4	04_5x10^-3_nM.txt	5×10^{-3} nM	1 200
5	05_0.05_nM.txt	0.05 nM	1 200
6	06_0.5_nM.txt	0.5 nM	1 200
7	07_5_nM.txt	5 nM	1 200
8	08_50_nM.txt	50 nM	1 200
9	09_0.5_uM.txt	0.5 μM	1 200
10	10_5_uM.txt	5 μM	1 200
11	11_50_uM.txt	50 μM	1 200
12	12_0.5_M.txt	0.5 M	1 200
13	13_5_M.txt	5 M	1 200

TABLE 2.4 – Classes de concentration pour l'Arsenic (As)

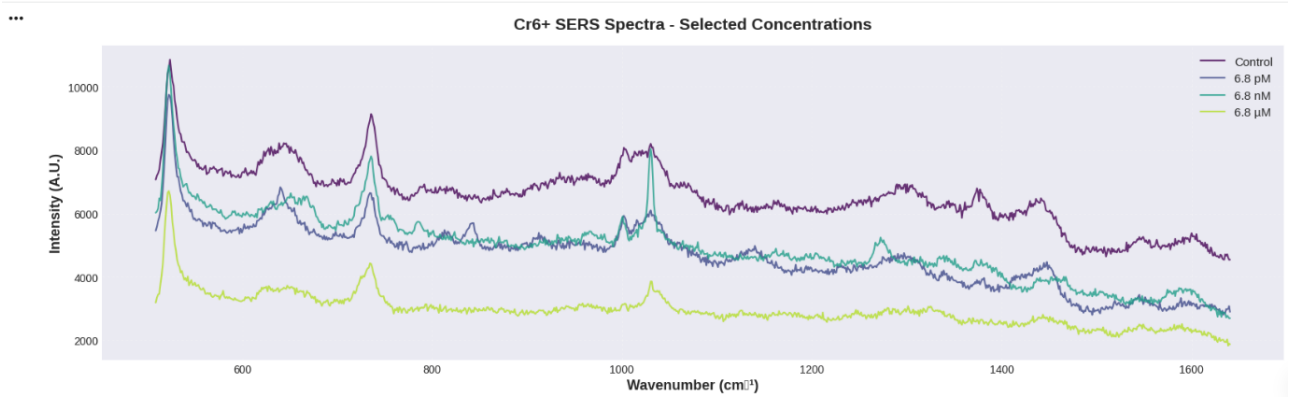

FIGURE 2.3 – Chrome hexavalent (Cr^{6+}) : Control, 68 pM, 68 nM et 68 ţM.

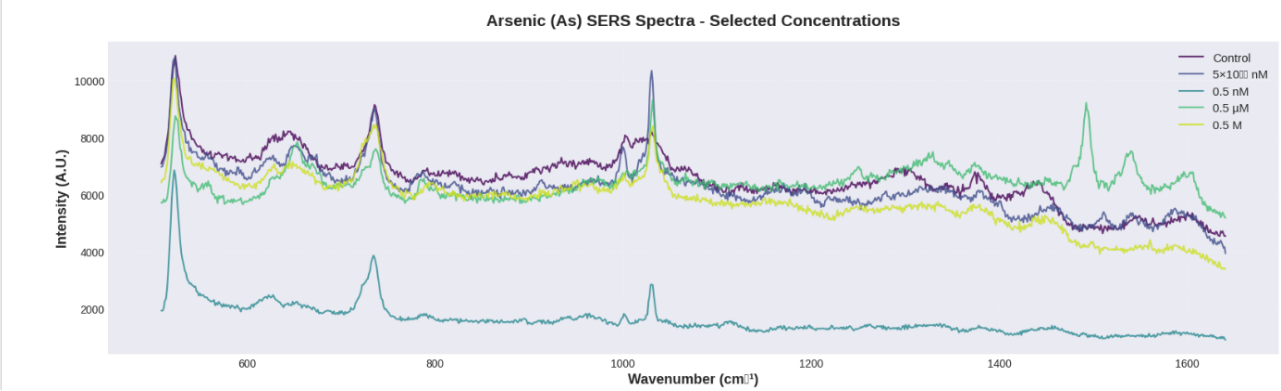

FIGURE 2.4 – Arsenic (As) : Control, 0.05 nM, 50 nM et 50 ţM.

FIGURE 2.5 – Exemples de spectres SERS pour différentes concentrations de Chrome et Arsenic

2.2.4 Résumé du dataset

Paramètre	**Chrome (Cr)**	**Arsenic (As)**	**Total**
Nombre total de spectres	18 800	23 600	42 400
Points par spectre	1 011	1 011	1 011
Nombre de classes	10	14	–
Plage de concentration	0.68 pM – 68 μM	5×10^{-3} pM – 5 M	–
Plage de nombres d'onde	508.88 – 1640.65 cm^{-1}	508.88 – 1640.65 cm^{-1}	–

TABLE 2.5 – Résumé global du dataset

Le **Control (témoin)** représente des échantillons d'eau non contaminée, servant de référence pour distinguer les spectres « propres » des spectres contenant des métaux lourds.

2.3 Prétraitement des Données

Le prétraitement des spectres SERS est une étape cruciale pour éliminer le bruit et les artefacts tout en préservant les informations discriminantes. Notre pipeline de prétraitement suit la méthodologie décrite dans l'article de référence.

2.3.1 Correction de la ligne de base (ALS)

La correction de la ligne de base est réalisée par la méthode *Asymmetric Least Squares* (ALS), permettant d'éliminer les contributions de fond et de fluorescence des spectres SERS.

2.3.1.0.1 Principe : L'algorithme ALS estime itérativement une ligne de base lisse qui suit les vallées du spectre tout en ignorant les pics. La fonction objective minimisée est définie comme suit :

$$\sum_i w_i(y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2 \tag{2.1}$$

Où :

— y_i : intensité du spectre original

— z_i : ligne de base estimée

— w_i : poids asymétrique

— λ : paramètre de lissage

— Δ^2 : opérateur de différence seconde

Paramètre	**Valeur**	**Description**
λ	10^6	Paramètre de lissage
p	0.01	Paramètre d'asymétrie
niter	10	Nombre d'itérations

TABLE 2.6 – Paramètres de la correction ALS

2.3.2 Lissage spectral (Savitzky–Golay)

Le filtre de Savitzky–Golay est appliqué afin de réduire le bruit haute fréquence tout en préservant les caractéristiques spectrales importantes telles que les pics et les formes spectrales.

2.3.2.0.1 Principe : Ce filtre repose sur une régression polynomiale locale appliquée sur une fenêtre glissante, où chaque point est remplacé par la valeur issue du polynôme ajusté.

Paramètre	**Valeur**	**Description**
Taille de la fenêtre	11 points	Nombre de points utilisés
Ordre du polynôme	3	Degré du polynôme d'ajustement

TABLE 2.7 – Paramètres du filtre Savitzky–Golay

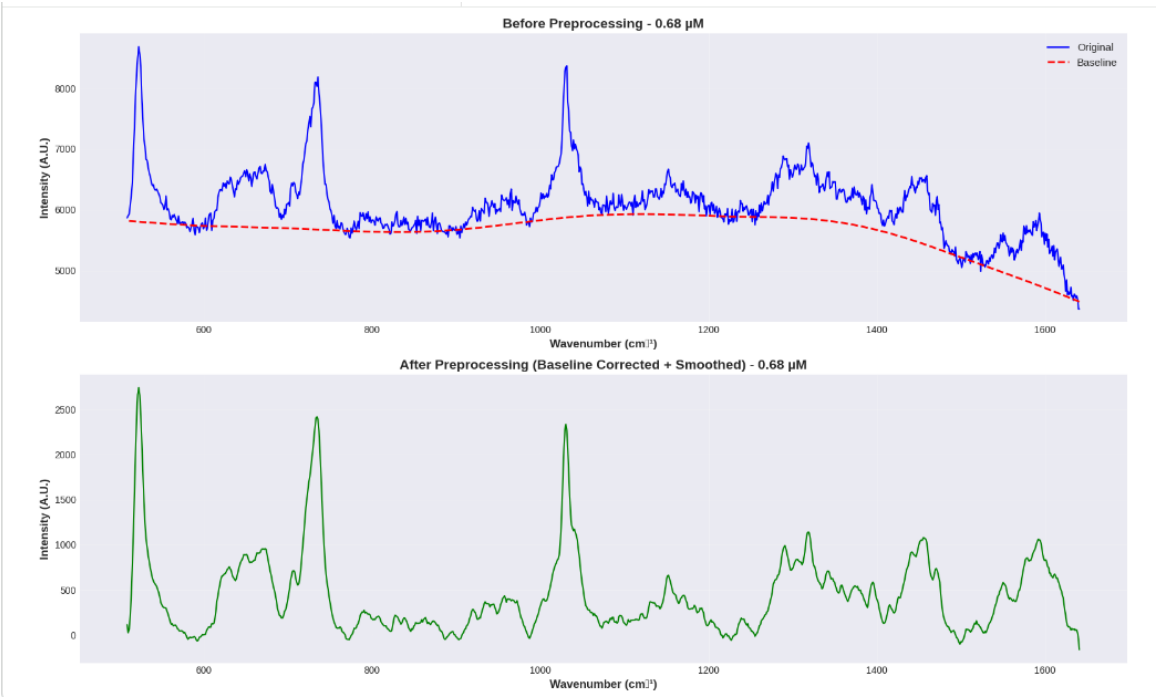

FIGURE 2.6 – Effet du prétraitement sur un spectre SERS
(a) Avant : spectre brut avec ligne de base (b) Après : spectre corrigé et lissé

2.3.3 Normalisation (StandardScaler)

La standardisation des données (Z-score) est appliquée à l'ensemble des spectres prétraités afin de centrer et réduire les variables :

$$Z = \frac{x - \mu}{\sigma} \tag{2.2}$$

Où :

— x : valeur originale

— μ : moyenne de la variable

— σ : écart-type de la variable

Cette étape garantit que toutes les variables possèdent une moyenne nulle et un écart-type unitaire, facilitant ainsi la convergence des algorithmes d'apprentissage automatique.

2.3.4 Réduction de dimensionnalité (PCA)

L'Analyse en Composantes Principales (PCA) est utilisée pour réduire la dimensionnalité des données spectrales tout en conservant l'information pertinente.

Paramètre	**Valeur**
Nombre de composantes	22
Dimensions originales	1011
Dimensions réduites	22
Variance expliquée	∼75%

TABLE 2.8 – Paramètres de la PCA

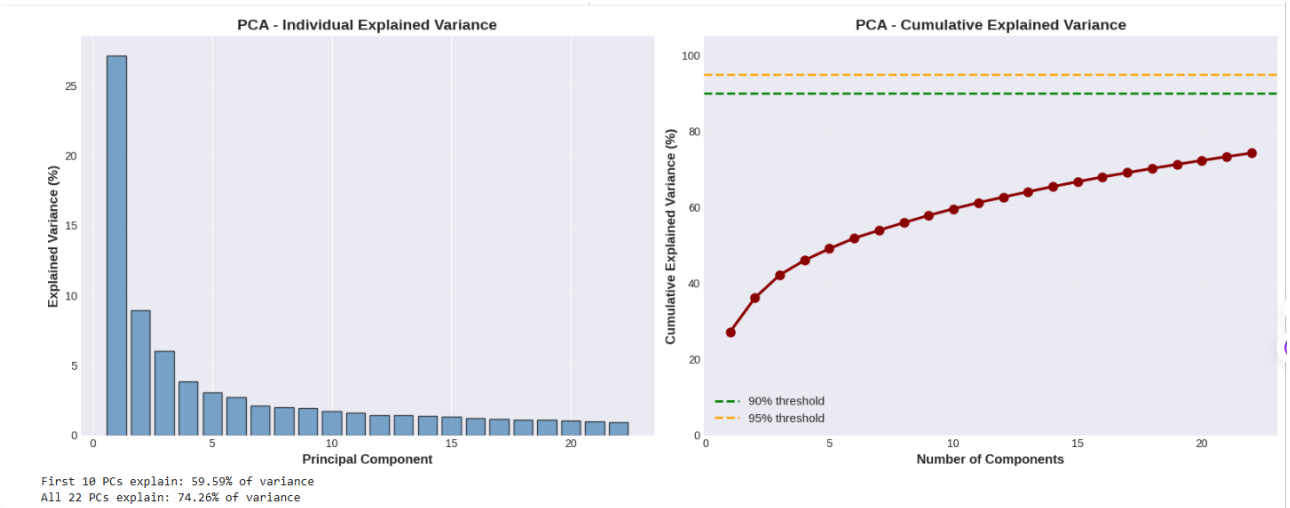

FIGURE 2.7 – Analyse de la variance expliquée par PCA

2.3.4.0.1 Avantages de la PCA :

— Réduction significative de la complexité computationnelle

— Élimination du bruit et de la redondance

— Conservation des caractéristiques spectrales discriminantes

— Amélioration de la généralisation des modèles

2.4 Division des Données et Équilibrage

2.4.1 Division Train/Test

Les données sont divisées selon le protocole standard de l'apprentissage automatique, comme présenté dans le tableau suivant :

Ensemble	**Proportion**	**Description**
Entraînement (Train)	80%	Apprentissage des modèles
Test	20%	Évaluation finale des performances

TABLE 2.9 – Répartition des ensembles d'apprentissage et de test

La division est réalisée de manière stratifiée afin de conserver la distribution proportionnelle des classes dans chaque ensemble, garantissant ainsi une représentation équitable de toutes les concentrations.

2.4.2 Équilibrage par SMOTE

Le jeu de données présente un déséquilibre entre les classes, notamment avec la classe *Control*, qui contient un nombre plus élevé d'échantillons. Afin de corriger ce déséquilibre, la technique *Synthetic Minority Over-sampling Technique* (SMOTE) est appliquée exclusivement à l'ensemble d'entraînement.

2.4.2.0.1 Principe de SMOTE : SMOTE génère des échantillons synthétiques pour les classes minoritaires en interpolant de nouveaux points à partir des échantillons existants et de leurs k plus proches voisins dans l'espace des caractéristiques.

2.4.2.0.2 Avantages :

— Équilibrage des classes sans perte d'information

— Amélioration des performances sur les classes minoritaires

— Réduction du biais en faveur des classes majoritaires

FIGURE 2.8 – SMOTE

2.4.2.0.3 Remarque importante : SMOTE est appliqué uniquement sur l'ensemble d'entraînement afin d'éviter toute fuite de données (*data leakage*) vers l'ensemble de test, ce qui garantirait une évaluation plus fiable des performances des modèles.

2.5 Outils et Technologies Utilisés

2.5.1 Langage de programmation

Le langage **Python 3.x** a été choisi comme langage principal pour ce projet en raison de ses nombreux avantages, résumés dans le tableau suivant :

Avantage	**Description**
Écosystème riche	Large collection de bibliothèques pour l'IA et le ML
Communauté active	Support et documentation abondants
Facilité d'utilisation	Syntaxe claire et lisible
Polyvalence	Adapté au ML, à la visualisation et au déploiement

TABLE 2.10 – Avantages du langage Python

2.5.2 Bibliothèques utilisées

Les principales bibliothèques Python utilisées dans ce projet sont présentées ici : .

— **Manipulation de données**

  — NumPy : Calculs numériques et tableaux

  — Pandas : Manipulation de DataFrames

— **Visualisation**

  — Matplotlib : Graphiques de base

  — Seaborn : Visualisations statistiques

— **Prétraitement**

  — SciPy : Filtre Savitzky–Golay, ALS

— **Machine Learning**

  — Scikit-learn : Modèles ML, PCA, métriques

  — Imbalanced-learn : SMOTE pour l'équilibrage

— **Visualisation dimensionnelle**

  — Scikit-learn (t-SNE) : Visualisation t-SNE

2.5.3 Environnement de développement

Les outils utilisés pour le développement et l'exécution du projet sont résumés dans le tableau suivant :

Outil	**Utilisation**
Google Colab	Exécution des notebooks avec accès gratuit au GPU
Jupyter Notebook	Développement interactif local
Visual Studio Code	Édition et gestion du code source

Table 2.11 – Environnement de développement

Conclusion

Ce chapitre a présenté la méthodologie complète adoptée pour la réalisation de notre système de détection des métaux lourds. Nous avons détaillé les données spectrales SERS utilisées, comprenant plus de 42 000 spectres répartis entre le Chrome et l'Arsenic avec leurs différentes concentrations.

Les étapes de prétraitement (correction de ligne de base ALS, lissage Savitzky-Golay, normalisation et PCA) ont été décrites en détail, ainsi que les stratégies de division des données et d'équilibrage par SMOTE. Les outils et technologies Python sélectionnés ont été justifiés en fonction des besoins spécifiques du projet.

Le chapitre suivant sera consacré à la conception de l'architecture hiérarchique du système de détection et à la présentation des résultats expérimentaux obtenus.

Chapitre 3

Conception et Modélisation

Introduction

Après avoir présenté dans le chapitre précédent la méthodologie adoptée, les données utilisées et les étapes de prétraitement, ce troisième chapitre est consacré à la conception et à la modélisation de notre système de détection des métaux lourds. Nous présenterons l'architecture hiérarchique proposée, la conception détaillée de chaque modèle, les algorithmes de classification sélectionnés ainsi que les métriques d'évaluation utilisées pour mesurer les performances du système.

3.1 Architecture du Système de Détection

3.1.1 Approche hiérarchique

La détection des métaux lourds dans l'eau agricole présente un défi de classification à deux niveaux : identifier d'abord le type de contamination, puis quantifier sa concentration. Pour répondre à cette problématique, nous avons conçu une architecture hiérarchique composée de trois modèles travaillant en cascade.

Principe de l'architecture :

L'approche hiérarchique décompose le problème complexe de détection en sous-problèmes plus simples et spécialisés :

Cette architecture s'inspire du raisonnement humain : on identifie d'abord la nature du problème avant de chercher à le quantifier.

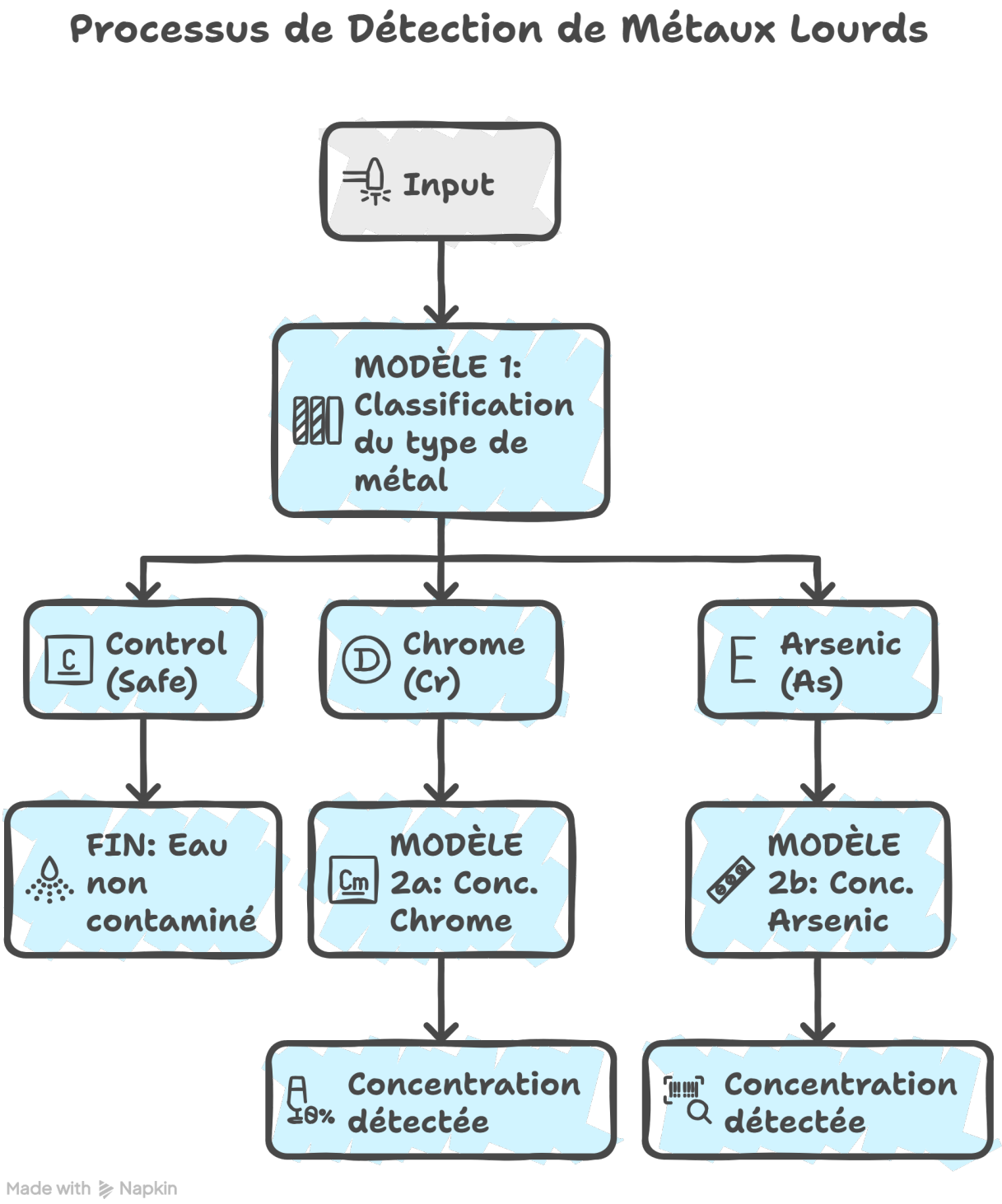

FIGURE 3.1 – Approche hiérarchique

3.1.2 Flux de prédiction

Le processus de prédiction suit un flux séquentiel bien défini :

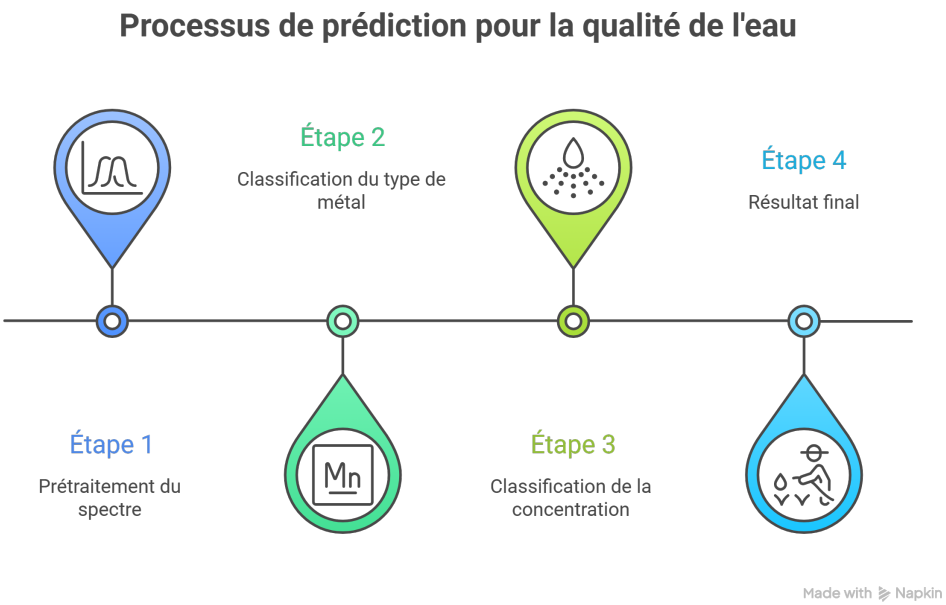

FIGURE 3.2 – Flux de prédiction détaillé

3.1.2.0.1 Étape 1 : Prétraitement du spectre

— Correction de la ligne de base (ALS)

— Lissage (Savitzky–Golay)

— Normalisation (StandardScaler)

— Réduction de dimensionnalité (PCA)

3.1.2.0.2 Étape 2 : Classification du type de métal (Modèle 1)

— **Entrée** : Spectre prétraité (22 composantes PCA)

— **Sortie** : Control, Chrome (Cr) ou Arsenic (As)

— Si **Control** → Fin (eau non contaminée)

— Si **Cr** ou **As** → Passage à létape 3

3.1.2.0.3 Étape 3 : Classification de la concentration (Modèle 2a ou 2b)

— Si Chrome détecté → Modèle 2a (9 classes de concentration)

— Si Arsenic détecté → Modèle 2b (13 classes de concentration)

3.1.2.0.4 Étape 4 : Résultat final

— Type de métal détecté

— Concentration estimée

— Statut de leau (Safe / Contaminée)

3.1.3 Avantages de larchitecture proposée

Larchitecture hiérarchique présente plusieurs avantages par rapport à une approche de classification directe :

— Spécialisation : Chaque modèle est optimisé pour une tâche spécifique

— Modularité : Les modèles peuvent être mis à jour indépendamment

— Interprétabilité : Processus de décision transparent et explicable

— Extensibilité : Ajout facile de nouveaux métaux (nouveau Modèle 2x)

— Performance : Meilleure précision grâce à la spécialisation

— Efficacité : Pas dexécution inutile si Control détecté

3.2 Conception des Modèles

3.2.1 Modèle 1 : Classification du type de métal

Objectif : Classifier les spectres SERS en trois catégories : Control, Chrome (Cr) ou Arsenic (As).

Spécifications du Modèle 1

Type de tâche : Classification multiclasse

Nombre de classes : 3 (Control, Cr, As)

Entrée : Vecteur de 22 composantes PCA

Sortie : Classe prédite (0, 1 ou 2)

Composition des données dentraînement

Classe 0 (Control) : Eau non contaminée, environ 4 000 spectres

Classe 1 (Cr) : Chrome (toutes concentrations), environ 10 800 spectres

Classe 2 (As) : Arsenic (toutes concentrations), environ 15 600 spectres

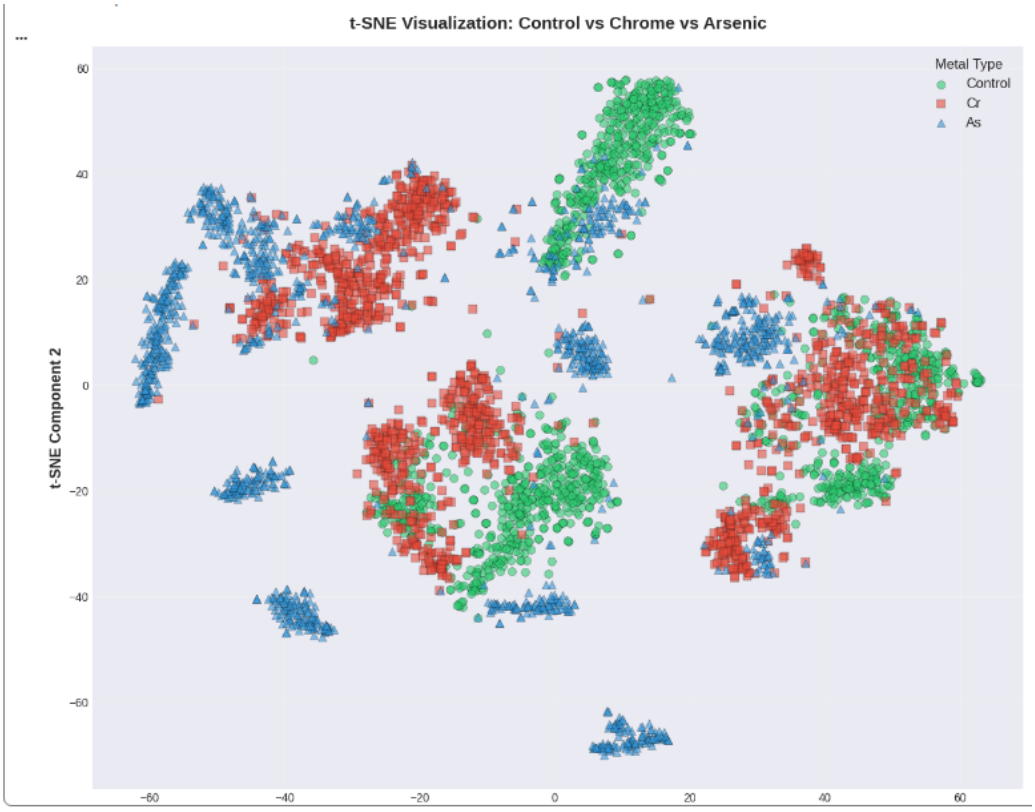

FIGURE 3.3 – Visualisation t-SNE des classes

Rôle critique : Ce modèle constitue le point dentrée du système. Toute erreur à ce niveau se propage aux étapes suivantes.

3.2.2 Modèle 2a : Classification de la concentration (Chrome)

Objectif : Prédire la concentration de Chrome parmi 9 niveaux distincts.

Spécifications du Modèle 2a

Type de tâche : Classification multiclasse

Nombre de classes : 9 concentrations distinctes

Entrée : Vecteur de 22 composantes PCA

Sortie : Index de concentration compris entre 0 et 8

3.2.3 Modèle 2b : Classification de la concentration (Arsenic)

Objectif : Prédire la concentration dArsenic parmi 13 niveaux distincts.

Spécifications du Modèle 2b

Type de tâche : Classification multiclasse

Nombre de classes : 13 concentrations distinctes

Entrée : Vecteur de 22 composantes PCA

Sortie : Index de concentration compris entre 0 et 12

3.2.3.1 Résumé comparatif des modèles

Caractéristique	**Modèle 1**	**Modèle 2a**	**Modèle 2b**
Objectif	Type de métal	Conc. Chrome	Conc. Arsenic
Classes	3	9	13
Complexité	Faible	Moyenne	Élevée
Données	Cr + As + Control	Cr uniquement	As uniquement

TABLE 3.1 – Comparaison des trois modèles

3.3 Algorithmes de Classification

Pour chaque modèle, plusieurs algorithmes de Machine Learning ont été évalués afin de sélectionner le plus performant. Cette section présente les algorithmes utilisés et leurs caractéristiques.

3.3.1 Support Vector Machine (SVM)

Le *Support Vector Machine* est un algorithme de classification supervisée qui vise à déterminer lhyperplan optimal séparant les différentes classes en maximisant la marge.

Principe

— Projection des données dans un espace de dimension supérieure

— Recherche de lhyperplan maximisant la marge entre les classes

— Utilisation de fonctions noyau pour traiter les cas non linéaires

Configuration utilisée

Kernel : RBF (Radial Basis Function), noyau gaussien

C : 1.0, paramètre de régularisation

Gamma : `scale`, coefficient du noyau RBF

Avantages

— Efficace en haute dimension

— Robuste au surapprentissage grâce à la régularisation

— Performant sur les données spectrales

Inconvénients

— Temps dentraînement élevé sur de grands ensembles de données

— Sensible au choix des hyperparamètres

3.3.2 Random Forest

Le *Random Forest* est un algorithme densemble basé sur la combinaison de plusieurs arbres de décision entraînés sur des sous-échantillons aléatoires des données.

Principe

— Construction de N arbres de décision indépendants

— Entraînement de chaque arbre sur un échantillon bootstrap

— Prédiction finale obtenue par vote majoritaire

Configuration utilisée

n_estimators : 200, nombre darbres

max_depth : 25, profondeur maximale des arbres

random_state : 42, graine aléatoire pour la reproductibilité

Avantages

— Robuste au surapprentissage

— Adapté aux données de haute dimension

— Fournit une mesure de limportance des variables

— Nécessite peu de prétraitement

Inconvénients

— Modèle volumineux en mémoire

— Moins interprétable quun arbre de décision unique

3.3.3 K-Nearest Neighbors (KNN)

Le *K-Nearest Neighbors* est un algorithme non paramétrique qui classe un échantillon selon les classes de ses k plus proches voisins.

Principe

— Calcul de la distance entre léchantillon et les données dentraînement

— Sélection des k voisins les plus proches

— Attribution de la classe majoritaire

Configuration utilisée

n_neighbors : 5, nombre de voisins

metric : Distance euclidienne

Avantages

— Simple à comprendre et à implémenter

— Absence de phase dentraînement explicite

— Efficace lorsque les classes sont bien séparées

Inconvénients

— Coût de prédiction élevé

— Sensible au bruit

— Dépend fortement du choix de k

3.3.4 Gradient Boosting

Le *Gradient Boosting* est un algorithme densemble qui construit séquentiellement des arbres de décision, chaque nouvel arbre corrigeant les erreurs du précédent.

Principe

— Entraînement dun modèle initial simple

— Calcul des résidus (erreurs)

— Apprentissage itératif de modèles corrigeant ces erreurs

Configuration utilisée

n_estimators : 300, nombre ditérations

max_depth : 4, profondeur des arbres

learning_rate : 0.05, taux dapprentissage

subsample : 0.8, fraction déchantillons par itération

Avantages

— Très bonnes performances globales

— Gère efficacement les relations non linéaires

— Robuste aux valeurs aberrantes

Inconvénients

— Temps dentraînement élevé

— Risque de surapprentissage si mal configuré

— Nombre important dhyperparamètres

3.3.5 Justification des choix

Algorithme	**Complexité**	**Interprétabilité**	**Adapté aux spectres**	**Temps dentraînement**
SVM (RBF)	Moyenne	Faible	Excellent	Moyen
Random Forest	Moyenne	Moyenne	Très bon	Rapide
KNN	Faible	Élevée	Bon	Très rapide
Gradient Boosting	Élevée	Faible	Excellent	Lent
Régression Logistique	Faible	Élevée	Limité	Très rapide

TABLE 3.2 – Comparaison des algorithmes de classification

Pour notre système, les algorithmes **SVM** et **Random Forest** ont été privilégiés car :

— ils sont particulièrement efficaces sur les données spectrales,

— ils gèrent efficacement les espaces de haute dimension,

— ils offrent un bon compromis entre performance et temps de calcul.

3.4 Métriques d'Évaluation

Lévaluation des performances des modèles nécessite lutilisation de métriques adaptées aux problèmes de classification multiclasse.

3.4.1 Accuracy (Exactitude)

L*Accuracy* représente la proportion de prédictions correctes parmi lensemble des prédictions effectuées.

$$Accuracy = \frac{Nombre\ de\ prdictions\ correctes}{Nombre\ total\ de\ prdictions} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Interprétation

— Métrique simple et intuitive

— Peut être trompeuse en présence de classes déséquilibrées

— Utile pour une évaluation globale rapide

3.4.2 Precision et Recall

Precision (Précision) La précision mesure la proportion de vrais positifs parmi les prédictions positives.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

Interprétation : Parmi les échantillons prédits comme positifs, combien le sont réellement ?

Recall (Rappel ou Sensibilité) Le rappel mesure la proportion de vrais positifs parmi les échantillons réellement positifs.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

Interprétation : Parmi les échantillons réellement positifs, combien ont été correctement détectés ?

3.4.3 F1-Score

Le *F1-Score* est la moyenne harmonique de la précision et du rappel, permettant dobtenir un compromis entre ces deux métriques.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.4}$$

Variantes pour le multiclasse

Macro F1 : Moyenne arithmétique des F1-scores de chaque classe

Weighted F1 : Moyenne pondérée par le nombre déchantillons par classe

Micro F1 : F1-score calculé globalement sur lensemble des classes

Dans cette étude, le **Weighted F1-Score** est privilégié afin de tenir compte du déséquilibre entre les classes.

3.4.4 Matrice de confusion

La matrice de confusion est un outil visuel permettant danalyser les performances dun modèle de classification en comparant les prédictions aux valeurs réelles.

Structure pour un problème à trois classes (Modèle 1)

	Prédit Control	Prédit Cr	Prédit As
Réel Control	Vrai Control	Erreur	Erreur
Réel Cr	Erreur	Vrai Cr	Erreur
Réel As	Erreur	Erreur	Vrai As

Interprétation

— La diagonale correspond aux prédictions correctes

— Les éléments hors diagonale représentent les erreurs de classification

— Permet didentifier les confusions entre classes spécifiques

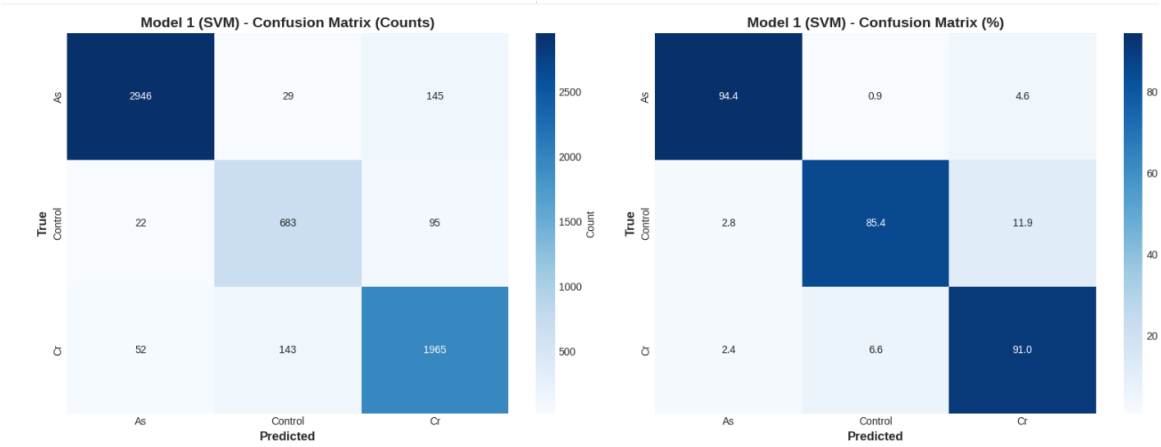

FIGURE 3.4 – Exemple de matrice de confusion

3.4.5 Validation croisée

La validation croisée *k-fold* est une technique dévaluation robuste consistant à diviser les données en k sous-ensembles.

Principe

— Division des données en k sous-ensembles (folds)

— Pour chaque fold i :

  — Entraînement sur les $k-1$ folds restants

  — Évaluation sur le fold i

— Calcul de la moyenne et de lécart-type des performances

3.4.6 Résumé des métriques utilisées

Accuracy : Évaluation globale des performances, simple et intuitive

Precision : Fiabilité des prédictions positives

Recall : Capacité à détecter les vrais positifs

F1-Score : Compromis entre précision et rappel

Matrice de confusion : Analyse détaillée et visuelle des erreurs

Validation croisée : Évaluation robuste et généralisable

Conclusion

Ce chapitre a présenté la conception et la modélisation de notre système de détection des métaux lourds basé sur une architecture hiérarchique. Cette approche décompose le problème en trois modèles spécialisés : un premier modèle pour identifier le type de métal (Control, Chrome ou Arsenic), et deux modèles dédiés à la prédiction des concentrations.

Nous avons détaillé les cinq algorithmes de classification évalués (SVM, Random Forest, KNN, Gradient Boosting et Régression Logistique) avec leurs configurations respectives. Les métriques d'évaluation (Accuracy, F1-Score, matrice de confusion et validation croisée) ont été présentées pour permettre une analyse rigoureuse des performances.

Le chapitre suivant présentera les résultats expérimentaux obtenus avec cette architecture, incluant la comparaison des algorithmes et l'évaluation du système complet.

Chapitre 4

Résultats et Discussion

Introduction

Ce quatrième chapitre est consacré à la présentation et à l'analyse des résultats obtenus lors de l'implémentation de notre système de détection des métaux lourds dans l'eau agricole. Après avoir exposé dans les chapitres précédents le cadre théorique, la méthodologie adoptée et l'architecture de notre solution, nous présentons ici les performances concrètes de nos modèles d'apprentissage automatique.

4.1 Résultats du Modèle 1 : Classification du Type de Métal

Le Modèle 1 constitue la première étape du système hiérarchique. Il est chargé de classifier les spectres SERS en trois catégories distinctes : *Control*, *Chrome (Cr)* et *Arsenic (As)*.

4.1.1 Comparaison des algorithmes

Deux algorithmes de classification ont été évalués pour cette tâche : la machine à vecteurs de support (SVM) avec noyau RBF et lalgorithme Random Forest. Le tableau 4.1 présente les résultats comparatifs obtenus.

Algorithme	**Accuracy (%)**	**F1-Score**
SVM (RBF)	92.01	0.9209
Random Forest	91.71	0.9170

Table 4.1 – Comparaison des algorithmes pour le Modèle 1

4.1.2 Matrice de confusion

La figure 4.2 présente la matrice de confusion du meilleur modèle sélectionné pour la classification du type de métal.

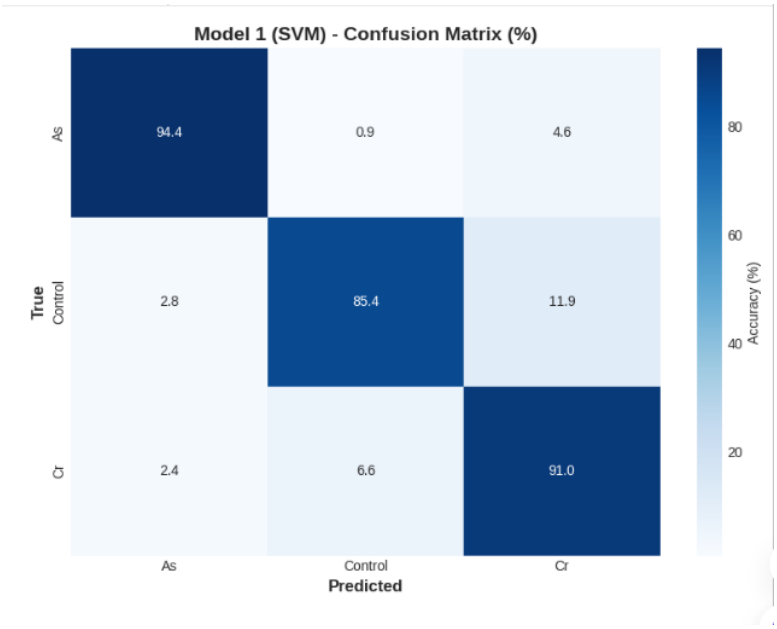

FIGURE 4.1 – Matrice de confusion du Modèle 1 (Classification du type de métal)

Lanalyse de la matrice de confusion met en évidence :

— Une excellente discrimination entre les échantillons *Control* et les échantillons contaminés ;

— Une bonne séparation entre les classes *Chrome* et *Arsenic* ;

— Des erreurs de classification principalement liées aux similarités spectrales observées à certaines concentrations.

4.1.3 Métriques de performance détaillées

Le figure 4.2 présente les métriques de performance détaillées pour chaque classe.

Classification Report - Model 1:

	precision	recall	f1-score	support
As	0.9755	0.9442	0.9596	3120
Control	0.7988	0.8538	0.8254	800
Cr	0.8912	0.9097	0.9003	2160
accuracy			0.9201	6080
macro avg	0.8885	0.9026	0.8951	6080
weighted avg	0.9223	0.9201	0.9209	6080

FIGURE 4.2 – Métriques par classe pour le Modèle 1

4.2 Résultats du Modèle 2a : Concentration du Chrome

Le Modèle 2a est activé lorsque le Modèle 1 détecte la présence de Chrome. Il classifie le spectre parmi 9 niveaux de concentration.

4.2.1 Performance de classification

4 algorithmes de classification ont été évalués pour cette tâche : la machine à vecteurs de support (SVM) avec noyau RBF et lalgorithme Random Forest, KNN, Gradient Boosting. Le tableau 4.2 présente les résultats comparatifs obtenus.

Algorithme	**Accuracy (%)**	**F1-Score**
SVM (RBF)	90.60	0.9058
Random Forest	83.67	0.8402
Gradient Boosting	83.29	0.8360
KNN	83.90	0.8417

TABLE 4.2 – Résultats du Modèle 2a (Classification de la concentration du Chrome)

4.2.1.1 Analyse des erreurs de prédiction

Lanalyse des erreurs de classification met en évidence les observations suivantes :

— Les confusions se produisent principalement entre des concentrations adjacentes ;

— Les concentrations extrêmes (très faibles et très élevées) sont mieux classifiées ;

— Les concentrations intermédiaires présentent un chevauchement spectral plus important, rendant leur discrimination plus difficile.

4.2.2 Matrice de confusion

La figure 4.3 illustre la matrice de confusion du Modèle 2a pour la classification des concentrations de Chrome.

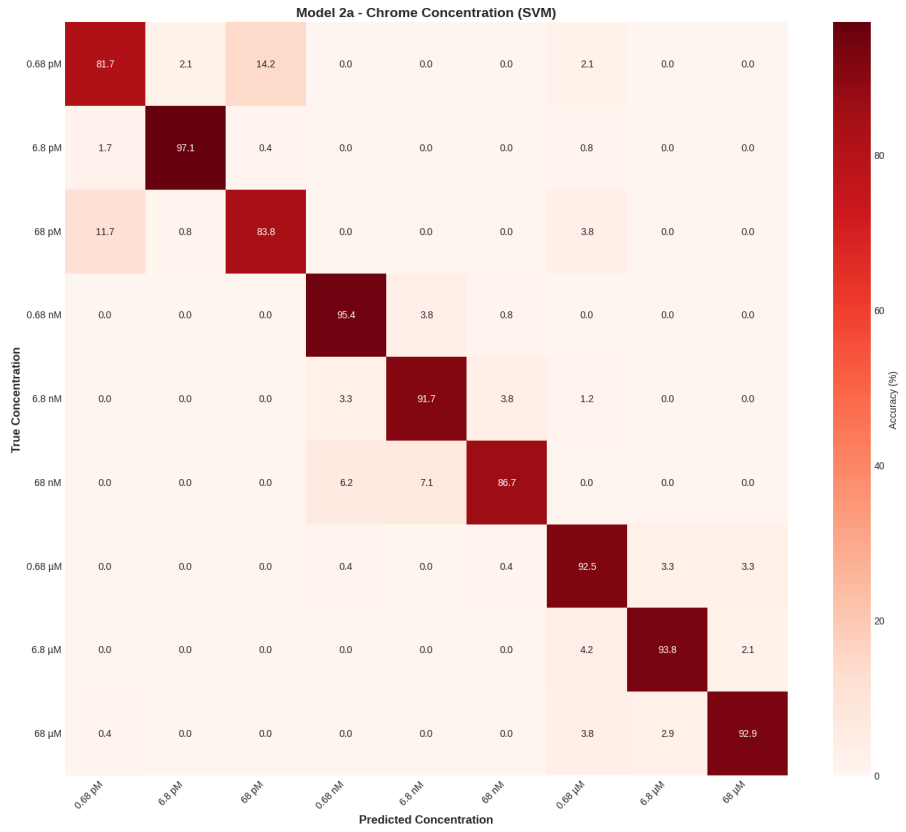

FIGURE 4.3 – Matrice de confusion du Modèle 2a (Concentration du Chrome)

La matrice de confusion normalisée montre une diagonale dominante, indiquant une bonne performance globale du modèle et une majorité de prédictions correctes pour chaque classe de concentration.

4.3 Résultats du Modèle 2b : Classification de la Concentration de lArsenic

Le Modèle 2b est appliqué aux spectres identifiés comme contenant de lArsenic (As). Il vise à prédire le niveau de concentration parmi treize classes distinctes, ce qui rend la tâche plus complexe que pour le Chrome.

4.3.1 Performance de classification

Les performances des deux algorithmes évalués sont présentées dans le tableau 4.3.

Algorithme	**Accuracy (%)**	**F1-Score**
SVM (RBF)	94.23	0.9061
Random Forest	90.22	0.9025
Gradient Boosting	89.96	0.9009
KNN	90.30	0.9048

TABLE 4.3 – Résultats du Modèle 2b (Classification de la concentration de lArsenic)

4.3.1.1 Analyse des erreurs de prédiction

Avec treize classes de concentration, la classification de lArsenic représente un défi plus important. Les principales observations sont les suivantes :

— La classification reste globalement performante malgré le nombre élevé de classes ;

— Les erreurs suivent un schéma similaire à celui observé pour le Chrome, avec des confusions entre classes adjacentes ;

— Les concentrations situées aux extrémités de la gamme sont les mieux reconnues par le modèle.

4.3.1.2 Matrice de confusion

La figure 4.4 présente la matrice de confusion du Modèle 2b pour la classification des concentrations dArsenic.

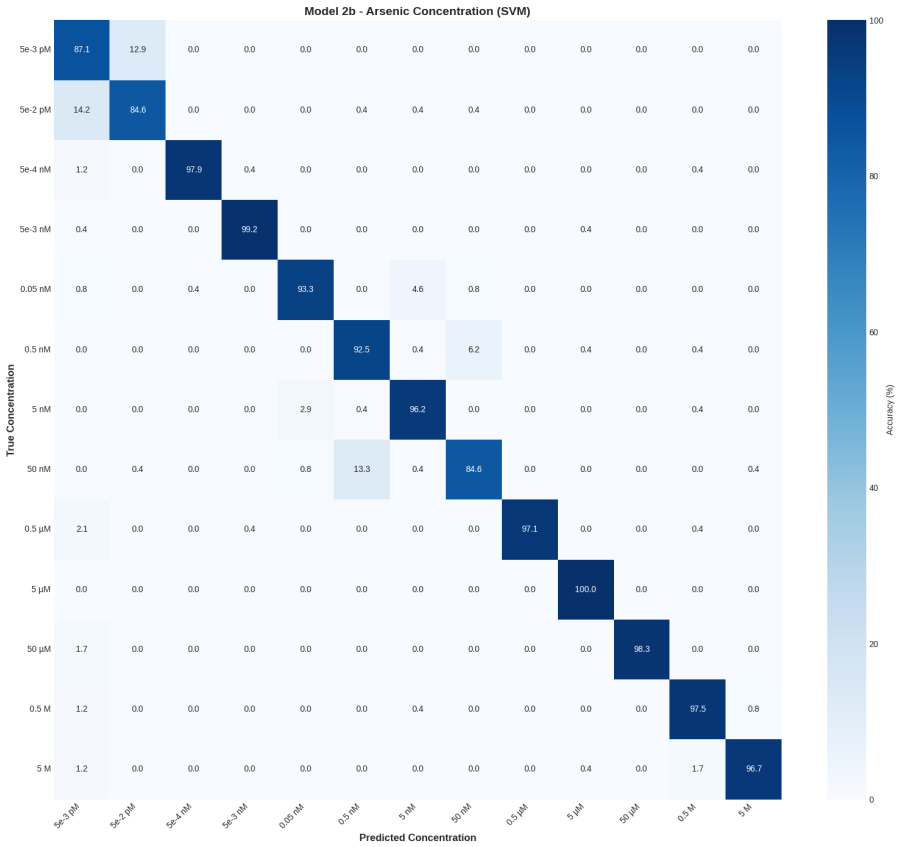

FIGURE 4.4 – Matrice de confusion du Modèle 2b (Concentration de lArsenic)

4.4 Évaluation du Système Hiérarchique Complet

4.4.1 Test du pipeline de bout en bout

Le système hiérarchique complet a été évalué de manière *end-to-end* sur un ensemble de données de test contenant des échantillons issus des trois catégories (Control, Chrome et Arsenic). Les performances globales obtenues sont résumées ci-dessous :

Accuracy du Modèle 1 (Type de métal) 92.01%

Accuracy du Modèle 2a (Concentration du Chrome) 90.60%

Accuracy du Modèle 2b (Concentration de lArsenic) 94.23%

Précision de détection de la contamination 90.00%

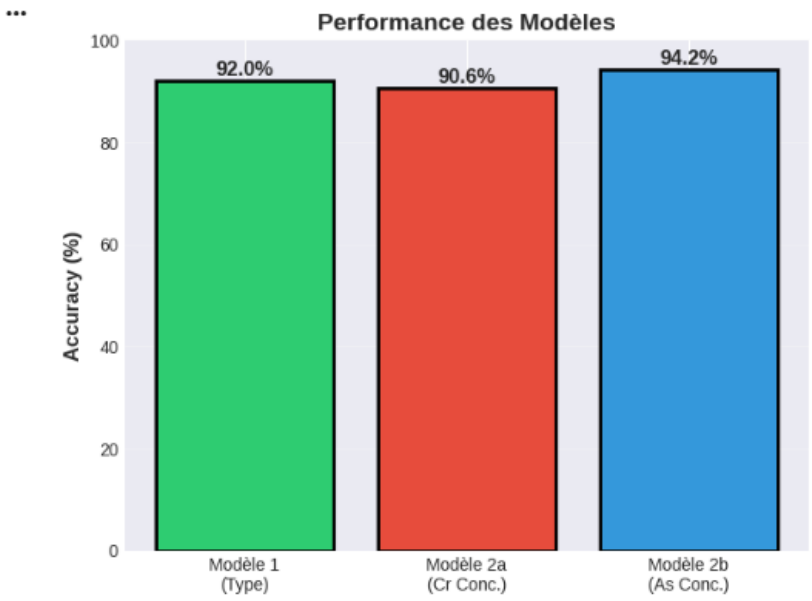

FIGURE 4.5 – Performance des modèles

Ces résultats montrent que le pipeline hiérarchique conserve de bonnes performances à chaque niveau de décision, depuis la détection de la contamination jusquà lestimation de la concentration.

4.4.2 Robustesse du système

Lévaluation de la robustesse du système met en évidence les points suivants :

— Le système maintient des performances stables sur différents ensembles déchantillons ;

— La validation croisée à 5 plis confirme la bonne capacité de généralisation des modèles ;

— Lutilisation de la méthode SMOTE pour léquilibrage des classes améliore significativement les performances sur les classes minoritaires.

4.5 Discussion

4.5.1 Interprétation des résultats

Les résultats obtenus démontrent lefficacité de lapproche hiérarchique proposée pour la détection et la quantification des métaux lourds dans leau.

Points forts

— La séparation entre échantillons *Control* et *Contaminés* est très fiable, garantissant une détection sûre de la contamination ;

— Larchitecture hiérarchique permet une spécialisation des modèles, ce qui améliore la précision globale du système ;

— Le prétraitement des spectres (ALS + filtre de Savitzky–Golay) améliore significativement la qualité des données dentrée.

Observations clés

— Les spectres SERS présentent des signatures distinctes pour chaque métal étudié ;

— La réduction de dimension par PCA à 22 composantes permet de conserver linformation pertinente tout en réduisant le bruit ;

— Le SVM avec noyau RBF se révèle particulièrement adapté à la classification de données spectrales.

4.5.2 Limites et perspectives damélioration

Limites identifiées

— Le système est actuellement limité à deux métaux (Chrome et Arsenic) ;

— Les cas de contamination multi-métaux ne sont pas encore pris en compte ;

— La généralisation à dautres sources deau nécessite une validation expérimentale supplémentaire.

Perspectives damélioration

— Extension à dautres métaux lourds : Plomb (Pb), Cadmium (Cd), Mercure (Hg) ;

— Détection multi-métaux avec gestion des contaminations simultanées ;

— Intégration dune classification *Safe / Unsafe* basée sur les seuils réglementaires (normes OMS) ;

— Développement dune interface utilisateur via *Streamlit* ;

— Exploration de modèles de Deep Learning, notamment des CNN 1D pour lextraction automatique de caractéristiques.

Conclusion

Ce chapitre a présenté les résultats expérimentaux du système hiérarchique de détection des métaux lourds basé sur la spectroscopie SERS et le Machine Learning. Les principales conclusions sont :

— Le Modèle 1 identifie efficacement le type de contamination (Control / Cr / As) avec une accuracy de 92.01% ;

— Le Modèle 2a prédit la concentration du Chrome parmi neuf niveaux avec une accuracy de 90.60% ;

— Le Modèle 2b estime la concentration de lArsenic parmi treize niveaux avec une accuracy de 94.23% ;

— Larchitecture hiérarchique permet une détection fiable et une quantification précise des métaux lourds dans leau agricole.

Ces résultats valident le potentiel de la combinaison *SERS + Machine Learning* pour la surveillance de la qualité de leau, ouvrant la voie à des applications concrètes en sécurité alimentaire et environnementale.

Conclusion Générale

La pollution des eaux agricoles par les métaux lourds représente un défi pour la santé publique et la sécurité alimentaire. Dans ce contexte, ce projet a porté sur le développement d'un système intelligent de détection des métaux lourds dans l'eau agricole, combinant la spectroscopie Raman exaltée de surface (SERS) et les techniques d'apprentissage automatique.

Au cours de ce projet, nous avons mené une étude bibliographique approfondie sur les métaux lourds et les méthodes de détection existantes. Nous avons ensuite conçu une architecture de classification hiérarchique à deux niveaux : un premier modèle pour identifier le type de contamination (Control, Chrome, Arsenic), puis des modèles spécialisés pour quantifier la concentration de chaque métal. Le système intègre un pipeline complet de traitement incluant la correction de ligne de base ALS, le lissage Savitzky-Golay, la réduction par PCA, l'équilibrage par SMOTE, et la classification par SVM et Random Forest.

Les expérimentations menées ont démontré l'efficacité de cette approche avec des performances satisfaisantes pour les trois modèles. Les contributions principales incluent un système de détection multi-métaux, une architecture hiérarchique innovante, et une méthodologie transférable à d'autres contaminants.

Certaines limites ont été identifiées, notamment la restriction à deux métaux et l'absence de traitement des contaminations simultanées. Les perspectives d'amélioration incluent l'extension à d'autres métaux lourds (Plomb, Cadmium, Mercure), l'ajout d'une classification Safe/Unsafe selon les normes OMS, le développement d'une interface utilisateur, et l'exploration des architectures Deep Learning.

En conclusion, ce projet a démontré le potentiel de la combinaison SERS et Machine Learning pour la détection rapide et précise des métaux lourds dans l'eau agricole, ouvrant des perspectives prometteuses pour la surveillance environnementale et la protection de la santé publique.

Bibliographie

[1] : Wei, S., Zhang, S., Su, M., Wang, X., Wang, J., Song, Y. (2023). Surface-enhanced Raman spectroscopy dataset for heavy metal detection in agricultural water using silver nanoparticles. Scientific Data, 10(1), 245. [lien]

[1] 2003 : Järup, L. *Hazards of heavy metal contamination*. British Medical Bulletin, 68(1), 167-182. [Lien]

[2] 2012 : Tchounwou, P. B., Yedjou, C. G., Patlolla, A. K., & Sutton, D. J. *Heavy metal toxicity and the environment*. Molecular, Clinical and Environmental Toxicology, 101, 133-164. [Lien]

[3] 2013 : Ali, H., Khan, E., & Sajad, M. A. *Phytoremediation of heavy metalsConcepts and applications*. Chemosphere, 91(7), 869-881. [Lien]

[4] 2010 : Nagajyoti, P. C., Lee, K. D., & Sreekanth, T. V. M. *Heavy metals, occurrence and toxicity for plants : a review*. Environmental Chemistry Letters, 8(3), 199-216. [Lien]

[5] 2017 : World Health Organization (WHO). *Guidelines for drinking-water quality : fourth edition incorporating the first addendum*. Geneva : WHO Press. [Lien]

[6] 1974 : Fleischmann, M., Hendra, P. J., & McQuillan, A. J. *Raman spectra of pyridine adsorbed at a silver electrode*. Chemical Physics Letters, 26(2), 163-166. [Lien]

[7] 2012 : Sharma, B., Frontiera, R. R., Henry, A. I., Ringe, E., & Van Duyne, R. P. *SERS : Materials, applications, and the future*. Materials Today, 15(1-2), 16-25. [Lien]

[8] 2020 : Langer, J., et al. *Present and future of surface-enhanced Raman scattering.* ACS Nano, 14(1), 28-117. [Lien]

[9] 2011 : Li, D., Qu, L., Zhai, W., Xue, J., Fossey, J. S., & Long, Y. *Facile on-site detection of substituted aromatic pollutants in water using thin layer chromatography combined with surface-enhanced Raman spectroscopy.* Environmental Science & Technology, 45(9), 4046-4052. [Lien]

[10] 2011 : Bantz, K. C., et al. *Recent progress in SERS biosensing.* Physical Chemistry Chemical Physics, 13(24), 11551-11567. [Lien]

[11] 2009 : Hastie, T., Tibshirani, R., & Friedman, J. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* Springer. [Lien]

[12] 2001 : Breiman, L. *Random forests.* Machine Learning, 45(1), 5-32. [Lien]

[13] 1995 : Cortes, C., & Vapnik, V. *Support-vector networks.* Machine Learning, 20(3), 273-297. [Lien]

[14] 2011 : Pedregosa, F., et al. *Scikit-learn : Machine learning in Python.* Journal of Machine Learning Research, 12, 2825-2830. [Lien]

[15] 2002 : Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. *SMOTE : Synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 16, 321-357. [Lien]

[16] 1964 : Savitzky, A., & Golay, M. J. *Smoothing and differentiation of data by simplified least squares procedures.* Analytical Chemistry, 36(8), 1627-1639. [Lien]

[17] 2003 : Eilers, P. H. *A perfect smoother.* Analytical Chemistry, 75(14), 3631-3636. [Lien]

[18] 2009 : Rinnan, Å., Van Den Berg, F., & Engelsen, S. B. *Review of the most common*

pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry, 28(10), 1201-1222. [Lien]