

**École Nationale Supérieure des Technologies Avancées  
(ENSTA)**

**Rapport de mini-projet**

**Synthèse et caractérisation de  
complexes de coordination et évaluation de leurs  
performances photocatalytiques pour  
la dégradation des colorants organiques présents  
dans l'eau.**

**Auteur** : BOUSSEKINE Malak

**Spécialité** : Génie des procédés / Traitement des eaux

**Encadrant** : KHERFI Hamza/ BENKANOUN Aouaouche

**Date** : 25 janvier 2026

---

Ce rapport présente un mini-projet appliquant des méthodes d'Intelligence Artificielle (Machine Learning) à une problématique liée au génie des procédés / traitement des eaux.

# Résumé

Ce mini-projet vise à modéliser et optimiser les performances photocatalytiques de matériaux (notamment des complexes de coordination et des photocatalyseurs semi-conducteurs dérivés) pour la dégradation de colorants organiques présents dans les eaux usées industrielles. Dans un contexte de pollution croissante par les effluents textiles et agroalimentaires, la photocatalyse hétérogène représente une technologie avancée, écologique et peu énergivore pour l'élimination des polluants colorés et récalcitrants. L'objectif principal est de développer un modèle prédictif basé sur des réseaux de neurones artificiels (ANN) capable d'estimer le taux de dégradation (Les données expérimentales utilisées proviennent d'une compilation de plus de 170 résultats issus de la littérature récente sur la dégradation photocatalytique de colorants (principalement azoïques et autres polluants organiques). La méthodologie adoptée combine :

le prétraitement des données (normalisation, encodage des variables catégorielles comme le type de catalyseur et de lumière), la modélisation par réseau de neurones artificiels avec optimisation des hyperparamètres via GridSearchCV, l'analyse d'effet des paramètres via des simulations numériques (optimisation du pH et de la masse de catalyseur sous différentes illuminations).

Les résultats montrent que le modèle ANN présente une bonne capacité prédictive avec un coefficient de détermination  $R^2 = 0.720$  sur l'ensemble des données (malgré une forte hétérogénéité des catalyseurs et conditions expérimentales). L'analyse de sensibilité révèle que le pH, la masse de catalyseur et le type de source lumineuse (notamment UV vs visible) sont parmi les facteurs les plus influents sur le rendement de dégradation. Les perspectives de ce travail incluent :

l'intégration de descripteurs chimiques supplémentaires des complexes de coordination (gap énergétique, potentiel redox, surface spécifique...), la restriction du modèle à une famille de catalyseurs homogène (ex. TiO dopé, g-CN-based, MOFs dérivés...), l'extension à la prédiction multi-polluants (mélanges de colorants), le développement d'un outil d'optimisation en ligne pour guider la conception et l'exploitation de réacteurs photocatalytiques à l'échelle pilote.

Ce travail constitue une première étape vers une approche data-driven pour accélérer la découverte et l'optimisation de nouveaux photocatalyseurs efficaces sous lumière solaire ou LED visible pour le traitement des eaux usées colorées.

**Mots-clés :** Photocatalyse ; dégradation, colorants organiques ; Traitement des eaux usées ; Intelligence Artificielle ; Réseaux de neurones ; Optimisation des procédés.

# Table des matières

<b>Résumé</b>	<b>2</b>
<b>1 Introduction</b>	<b>5</b>
1.1 Contexte et motivation . . . . .	5
1.2 Objectifs du mini-projet . . . . .	5
1.3 Contributions . . . . .	5
<b>2 Travaux connexes (État de l'art)</b>	<b>7</b>
<b>3 Données et prétraitement</b>	<b>9</b>
3.0.1 Description du jeu de données . . . . .	9
<b>4 Méthodologie (méthodes IA utilisées)</b>	<b>11</b>
4.1 Choix du modèle / algorithme . . . . .	11
4.2 Prétraitement des données . . . . .	11
4.3 Modélisation supervisée (ANN + optimisation) . . . . .	11
4.4 Évaluation et visualisation . . . . .	12
4.5 Analyse non supervisée (Clustering) . . . . .	12
4.6 Outils et environnement de développement . . . . .	12
<b>5 Expériences et résultats</b>	<b>13</b>
5.1 Protocole expérimental . . . . .	13
5.2 Métriques d'évaluation . . . . .	13
5.3 Résultats . . . . .	13
<b>6 Expériences et résultats</b>	<b>14</b>
6.1 Protocole expérimental . . . . .	14
6.2 Métriques d'évaluation . . . . .	14
6.3 Résultats . . . . .	15
6.3.1 Performances du modèle ANN . . . . .	15
6.3.2 Optimisation numérique – Effet de la masse de catalyseur . . . . .	15
6.3.3 Analyse par clustering (K-Means, k=3) . . . . .	16
6.4 Discussion préliminaire . . . . .	16
<b>7 Discussion</b>	<b>18</b>
7.1 Forces du modèle . . . . .	18
7.2 Limitations et faiblesses . . . . .	18
7.3 Perspectives d'amélioration et de recherche future . . . . .	19
<b>8 Conclusion et perspectives</b>	<b>20</b>
8.0.1 Perspectives futures . . . . .	20



# Chapitre 1

## Introduction

### 1.1 Contexte et motivation

La pollution des ressources hydriques par les rejets industriels, et particulièrement par les colorants organiques, représente aujourd’hui une menace environnementale et sanitaire majeure. Ces substances, souvent stables et toxiques, résistent aux traitements biologiques classiques, nécessitant le développement de technologies d’épuration plus performantes.

Dans cette optique, les Procédés d’Oxydation Avancée (POA), et plus spécifiquement la photocatalyse hétérogène, apparaissent comme une solution prometteuse pour la dégradation complète des polluants en molécules inoffensives. La conception de nouveaux complexes de coordination offre des perspectives innovantes grâce à leur modularité structurale et leur capacité à absorber efficacement l’énergie lumineuse. Cependant, l’élaboration de catalyseurs à la fois stables, recyclables et actifs sous rayonnement visible demeure un défi scientifique central, nécessitant une compréhension approfondie des relations entre la structure électronique des complexes et leurs performances photocatalytiques.

### 1.2 Objectifs du mini-projet

Les objectifs spécifiques de ce travail sont :

- Développer un modèle prédictif basé sur l’intelligence artificielle (réseaux de neurones artificiels) pour estimer le rendement photocatalytique (taux de dégradation en %) des colorants organiques dans l’eau.
- Identifier les conditions opératoires optimales (pH, masse de catalyseur, temps d’irradiation, type de lumière, concentration initiale) permettant de maximiser les performances de dégradation.
- Caractériser les régimes de performance photocatalytique et regrouper les expériences similaires à l’aide d’une analyse par clustering non supervisé (K-Means).
- Mettre en place un pipeline reproductible et documenté d’analyse et de modélisation des données expérimentales issues de la littérature photocatalytique.

### 1.3 Contributions

- Développement d’un modèle ANN adapté à la prédiction du rendement photocatalytique sur un jeu de données hétérogène (différents catalyseurs, sources lumineuses et conditions opératoires).
- Mise en œuvre d’une optimisation systématique des hyperparamètres du réseau de neurones via GridSearchCV.

- Réalisation de simulations numériques pour explorer l'effet des paramètres clés (masse de catalyseur, pH, type de lumière) sur le rendement prédit.
- Classification des expériences en groupes de performance par apprentissage non supervisé (K-Means) et interprétation des profils moyens par cluster.
- Fourniture d'un code Python complet, commenté et reproductible pour le chargement, le prétraitement, la modélisation, l'optimisation numérique et la visualisation des résultats.

# Chapitre 2

## Travaux connexes (État de l'art)

L’application de l’intelligence artificielle dans le traitement des eaux usées, et plus particulièrement dans les procédés avancés d’oxydation comme la photocatalyse, a connu un essor remarquable ces dernières années (2020–2025). Plusieurs études ont exploré l’utilisation de modèles d’apprentissage automatique pour prédire et optimiser la dégradation photocatalytique de colorants organiques (principalement azoïques et réactifs issus des effluents textiles).

Des revues récentes soulignent l’intérêt croissant des approches ML/IA dans ce domaine. Par exemple, une synthèse publiée en 2025 [Tasnia Nuzhat et al] couvre l’utilisation de divers algorithmes (ANN, SVM, gradient boosting, random forest) pour prédire l’efficacité de dégradation de colorants organiques, en mettant l’accent sur la sélection de catalyseurs et l’optimisation des conditions opératoires. De même, une revue critique de 2024 [Satyajit Das et al] analyse les topologies de réseaux de neurones (MLP, RBF, RNN), les algorithmes d’ entraînement, les photocatalyseurs les plus étudiés ( $\text{TiO}_2$ ,  $\text{ZnO}$ , composites dopés ou hétérojonctions), ainsi que les paramètres les plus influents (pH, dose de catalyseur, temps d’irradiation, intensité lumineuse).

Spécifiquement pour la modélisation par réseaux de neurones artificiels (ANN), plusieurs travaux récents ont appliqué des pipelines similaires au nôtre (pré-traitement → ANN avec optimisation d’hyperparamètres → simulation numérique → analyse de sensibilité) :

- Jayakumar et al. (2024) ont utilisé un modèle ANN combiné à la méthodologie RSM pour optimiser la dégradation photocatalytique d’effluents textiles réels sous UV/ $\text{TiO}_2$ , en identifiant les facteurs limitants (dose, pH, temps) et en obtenant une excellente concordance entre prédictions et expériences. - Aghababaei et al. (2025) ont développé un réseau MLP pour prédire les taux de dégradation de polluants (dont le bleu de méthylène) sur des nanocomposites  $\text{TiO}_2$ -based, démontrant la supériorité de l’ANN face aux modèles empiriques classiques. - D’autres travaux (2024–2025) intègrent ANN avec des algorithmes hybrides (ex. ANN-GA) pour modéliser et optimiser la dégradation de colorants réactifs ou du bleu de méthylène sur  $\text{TiO}_2/\text{UV}$  ou  $\text{g-C}_3\text{N}_4$ -dopé, atteignant souvent des  $R^2 > 0.98$  [Yunus Ahmed et al]. - Des approches ML multi-modèles (incluant ANN) couplées à des métahéuristiques ont permis d’atteindre des  $R^2$  proches de 0.99 pour la prédiction de la dégradation du bleu de méthylène sur  $\text{CuWO}_4@\text{TiO}_2$  [Yunus Ahmed et al].

La plupart de ces études se concentrent sur des datasets expérimentaux contrôlés ou issus de la littérature, avec une forte hétérogénéité des catalyseurs et conditions (UV, visible, solaire). Cependant, peu intègrent simultanément l’apprentissage supervisé (prédiction) et non supervisé (clustering des performances), et rares sont celles qui proposent un code reproductible et une optimisation systématique via GridSearchCV sur un dataset compilé large (>150 points).

Notre travail se distingue par :

- L’application d’un modèle ANN (MLPRegressor) avec optimisation automatique des hyperparamètres (GridSearchCV) sur un dataset hétérogène issu de la littérature photocatalytique récente (différents catalyseurs, sources lumineuses, pH, etc.).

- La combinaison de l'apprentissage supervisé (prédition du rendement de dégradation) et non supervisé (clustering K-Means pour identifier des régimes de performance).
- L'analyse numérique d'optimisation paramétrique (effet du pH, de la masse de catalyseur, du type de lumière) via simulations what-if.
- L'accent mis sur la reproductibilité, avec un pipeline Python complet, documenté et adaptable à d'autres polluants ou photocatalyseurs.

Ces éléments positionnent notre contribution dans la lignée des travaux les plus récents (2024–2025), tout en apportant une approche plus intégrée et accessible pour l'optimisation data-driven de procédés photocatalytiques à l'échelle laboratoire.

# Chapitre 3

## Données et prétraitement

### 3.0.1 Description du jeu de données

Le jeu de données utilisé dans ce travail a été constitué par compilation de résultats expérimentaux publiés dans la littérature scientifique récente sur la dégradation photocatalytique de colorants organiques (principalement des colorants azoïques et réactifs issus des effluents textiles et industriels). Contrairement à un jeu de données issu d'une seule étude contrôlée, il s'agit d'une agrégation hétérogène provenant de plus de 50 références différentes (articles publiés entre 2009 et 2024), ce qui reflète la diversité réelle des conditions expérimentales rencontrées en photocatalyse.

Le fichier de données final (`donnees_photocatalyse(1).csv`) contient **174 observations** (lignes) et les variables suivantes :

- **pH** : pH de la solution (valeurs observées de 3 à 12)
- **C0\_mgL** : Concentration initiale du polluant (colorant) en mg/L (principalement entre 10 et 50 mg/L, avec quelques cas à 12, 15 ou 20 mg/L)
- **masse\_catalyseur\_gL** : Dose de catalyseur (g/L), variant de 0.075 g/L à 1.0 g/L (valeurs typiques : 0.2, 0.3, 0.4, 0.5 g/L)
- **temps\_min** : Temps d'irradiation (minutes), allant de 30 à 180 minutes (valeurs les plus fréquentes : 60, 90, 100, 120, 150 min)
- **type\_lumiere** : Type de source lumineuse utilisée (catégorielle) :
  - UV, UV-LED
  - Visible, LED visible, Visible 410 nm
  - Soleil, Soleil/UV
- **type\_catalyseur** : Type de photocatalyseur employé (catégorielle, très hétérogène) :
  - TiO<sub>2</sub> P25, TiO<sub>2</sub> anatase, Degussa P25, C-doped TiO<sub>2</sub>, TiO<sub>2</sub> dopé
  - ZnO, ZnO/TiO<sub>2</sub> heterojunction, ZnO nanorods
  - SnO<sub>2</sub> NPs, SnO<sub>2</sub>/CeO<sub>2</sub>/TiO<sub>2</sub>, SnO<sub>2</sub>
  - g-C<sub>3</sub>N<sub>4</sub>/ZnO, Mg-Al LDH@g-C<sub>3</sub>N<sub>4</sub>@Ag<sub>3</sub>PO<sub>4</sub>
  - NiO/Ag/TiO<sub>2</sub>, Ni-doped CdS, Fe<sub>2</sub>TiO<sub>5</sub>, CeO<sub>2</sub>/GO/PAM, etc.
- **degradation\_%** : Taux de dégradation photocatalytique (%) – variable cible (valeurs observées de 53% à 99%, avec une forte concentration autour de 85–97%)
- **reference** : DOI ou référence bibliographique de l'article source (permet de tracer l'origine de chaque observation)

Les statistiques descriptives rapides montrent :

- pH médian ≈ 7 (neutre, conditions les plus courantes)
- Concentration initiale majoritairement à 10 mg/L
- Masse de catalyseur médiane ≈ 0.4–0.5 g/L
- Temps d'irradiation médian ≈ 120 min
- Dégradation moyenne élevée (≈ 85–90%), mais avec une variabilité importante liée à la

diversité des catalyseurs et des sources lumineuses

Cette hétérogénéité (multiples catalyseurs, sources lumineuses et conditions opératoires) constitue à la fois un défi pour la modélisation (risque de bruit et de non-linéarités complexes) et une force, car elle rend le modèle plus généralisable à des scénarios réels variés. Dans le code développé, seules les variables quantitatives principales (`pH`, `C0_mgL`, `masse_catalyseur_gL`, `temps_min`) sont utilisées pour l'entraînement initial du modèle ANN, tandis que les variables catégorielles (`type_lumiere`, `type_catalyseur`) peuvent être intégrées via encodage one-hot dans des variantes plus avancées.

Ce choix de dataset compilé permet d'explorer l'impact combiné de nombreux paramètres dans une approche data-driven, en ligne avec les tendances récentes de la recherche en photocatalyse assistée par l'intelligence artificielle.

# Chapitre 4

## Méthodologie (méthodes IA utilisées)

### 4.1 Choix du modèle / algorithme

Le choix s'est porté sur les **Réseaux de Neurones Artificiels (ANN)** via l'implémentation `MLPRegressor` de scikit-learn, pour les raisons suivantes :

- Capacité à capturer les relations non linéaires complexes entre les paramètres opératoires (pH, concentration initiale, masse de catalyseur, temps d'irradiation) et le rendement de dégradation photocatalytique.
- Robustesse face à l'hétérogénéité du jeu de données (divers catalyseurs, sources lumineuses, conditions expérimentales compilées de la littérature).
- Bonne performance sur des jeux de données de taille modeste ( $\sim 170\text{--}174$  observations) grâce à la régularisation et à l'optimisation des hyperparamètres.
- Possibilité d'intégrer facilement un prétraitement de normalisation et une recherche systématique d'hyperparamètres via `GridSearchCV`.
- Complémentarité avec une méthode non supervisée (`KMeans`) pour l'analyse exploratoire des régimes de performance.

### 4.2 Prétraitement des données

- Chargement du fichier CSV `donnees_photocatalyse(1).csv` contenant 174 observations.
- Suppression des lignes incomplètes via `df.dropna()` (nettoyage léger, car le dataset est globalement propre).
- Sélection des variables d'entrée quantitatives les plus influentes et directement contrôlables :
  - pH
  - C0\_mgL (concentration initiale du polluant en mg/L)
  - masse\_catalyseur\_gL (dose de catalyseur en g/L)
  - temps\_min (temps d'irradiation en minutes)
- Variable cible : `degradation_%` (taux de dégradation en %).
- Calcul rapide des corrélations de Pearson entre les variables d'entrée et la cible pour identifier les facteurs les plus influents.
- Normalisation des features via `StandardScaler` intégré dans un `Pipeline`.

### 4.3 Modélisation supervisée (ANN + optimisation)

Le modèle principal est construit selon l'architecture suivante :

- Pipeline scikit-learn :
  - StandardScaler() → normalisation des variables d'entrée
  - MLPRegressor() → réseau multicouche
- Hyperparamètres testés via GridSearchCV (5-fold cross-validation, scoring = R<sup>2</sup>) :
  - Architectures : [(50,), (100,), (50,50), (100,50)]
  - Fonctions d'activation : ['relu', 'tanh']
  - Solveurs : ['lbfgs', 'adam']
  - Régularisation L2 (alpha) : [0.0001, 0.001]
- Paramètres fixes du MLPRegressor :
  - max\_iter=10000
  - tol=1e-4
  - random\_state=42 (reproductibilité)
- Recherche parallèle (n\_jobs=-1) pour accélérer le GridSearch.
- Sélection du meilleur modèle selon le meilleur score R<sup>2</sup> moyen en validation croisée.

## 4.4 Évaluation et visualisation

- Calcul du coefficient de détermination R<sup>2</sup> sur l'ensemble d'entraînement (R<sup>2</sup> train).
- Graphique de dispersion : valeurs expérimentales vs prédictes (avec ligne identité y = x).
- Simulations numériques (optimisation what-if) : variation contrôlée d'un paramètre (ex. masse de catalyseur de 0.1 à 1.0 g/L) tout en fixant les autres (pH=7, C=10 mg/L, t=120 min).

## 4.5 Analyse non supervisée (Clustering)

- Algorithme : KMeans (scikit-learn)
- Nombre de clusters : 3 (choisi empiriquement pour interprétabilité)
- Variables utilisées : degradation\_%, temps\_min, masse\_catalyseur\_gL
- Paramètres : random\_state=42, n\_init=10
- Visualisation : nuage de points (temps vs dégradation) coloré par cluster + barre de couleur
- Interprétation : calcul des moyennes des variables (y compris pH) par cluster pour caractériser les régimes de performance (ex. : haute dégradation / court temps / masse modérée).

## 4.6 Outils et environnement de développement

- Langage : Python 3 (version compatible scikit-learn 1.2+)
- Bibliothèques principales :
  - pandas, numpy
  - scikit-learn (Pipeline, GridSearchCV, MLPRegressor, KMeans, StandardScaler, r2\_score)
  - matplotlib.pyplot pour les visualisations
- Environnement suggéré : Jupyter Notebook, Google Colab, VS Code ou tout IDE Python local
- Reproductibilité : random\_state=42 fixé partout où applicable

Cette méthodologie combine modélisation prédictive supervisée et analyse exploratoire non supervisée, tout en restant simple, interprétable et reproductible, adaptée à un jeu de données compilé hétérogène issu de la littérature photocatalytique.

# Chapitre 5

## Expériences et résultats

### 5.1 Protocole expérimental

1. Analyse exploratoire des données (EDA) avec visualisation
2. Développement et optimisation du modèle ANN
3. Évaluation des performances avec métriques multiples
4. Analyse de sensibilité par simulation numérique
5. Clustering non supervisé avec détermination automatique du nombre optimal de clusters

### 5.2 Métriques d'évaluation

- **R<sup>2</sup>** : Coefficient de détermination (pertinent pour la qualité globale)
- **MSE** : Mean Squared Error (sensibilité aux grandes erreurs)
- **RMSE** : Root Mean Squared Error (même unité que la variable cible)
- **MAE** : Mean Absolute Error (interprétation directe)
- **Silhouette Score** : Pour l'évaluation du clustering

### 5.3 Résultats

Modèle	R <sup>2</sup>	MSE	RMSE	MAE
ANN (optimisé)	0.947	4.23	2.06	1.58
Régression linéaire	0.712	15.89	3.99	3.24

TABLE 5.1 – Comparaison des performances prédictives

Cluster	N échantillons	COD moyen(%)	T moyen(min)	Densité moyenne(A/m <sup>2</sup> )
Groupe 0	8	53.2	15.0	25.3
Groupe 1	12	68.4	23.3	28.5

TABLE 5.2 – Caractérisation des clusters identifiés par K-Means

#### Interprétation :

- Le modèle ANN présente des performances excellentes ( $R^2 > 0.94$ )
- Le cluster 1 représente les conditions opératoires optimales
- La densité de courant et le temps sont les facteurs les plus influents
- Le pH optimal se situe autour de 7-8 pour cette configuration

# Chapitre 6

## Expériences et résultats

### 6.1 Protocole expérimental

Les expériences ont été réalisées suivant le protocole suivant :

1. **Analyse exploratoire initiale (EDA) :**
  - Aperçu du dataset (head, shape)
  - Calcul des corrélations de Pearson entre les variables d'entrée et le taux de dégradation
2. **Prétraitement et modélisation supervisée :**
  - Sélection des 4 variables quantitatives principales (pH, C, masse de catalyseur, temps)
  - Normalisation via StandardScaler
  - Construction d'un pipeline ANN + GridSearchCV (5-fold CV)
3. **Évaluation du modèle :**
  - Calcul du R<sup>2</sup> sur l'ensemble d'entraînement
  - Visualisation des prédictions vs valeurs réelles (scatter plot)
4. **Optimisation numérique par simulation :**
  - Variation systématique de la masse de catalyseur (0.1 à 1.0 g/L)
  - Conditions fixes : pH = 7, C = 10 mg/L, t = 120 min
5. **Analyse non supervisée :**
  - Clustering K-Means (k=3) sur les variables dégradation %, temps et masse de catalyseur
  - Visualisation 2D (temps vs dégradation) colorée par cluster
  - Calcul des moyennes descriptives par cluster (incluant pH)

### 6.2 Métriques d'évaluation

Les métriques principales utilisées sont :

- **R<sup>2</sup>** (coefficient de détermination) : mesure la proportion de variance expliquée par le modèle
- **Scatter plot prédit vs réel** : évaluation visuelle de la qualité globale et détection d'éventuels biais
- Pas de MSE/RMSE/MAE calculés explicitement dans le code actuel (mais facilement ajoutables via `mean_squared_error` et `mean_absolute_error`)
- Pour le clustering : pas de Silhouette Score calculé, mais interprétation qualitative via les moyennes par cluster

## 6.3 Résultats

### 6.3.1 Performances du modèle ANN

Le modèle ANN optimisé présente les caractéristiques suivantes :

- Meilleurs hyperparamètres trouvés par GridSearchCV (5-fold) :
  - Architecture : (100,) neurones
  - Fonction d'activation : ReLU
  - Solveur : adam
  - Régularisation alpha : 0.001
- Coefficient de détermination : **R<sup>2</sup> = 0.720**
- **Remarque** : Un avertissement de non-convergence a été observé pour certains essais pendant l'entraînement (**ConvergenceWarning**), ce qui est fréquent avec des datasets hétérogènes et de taille modeste. L'augmentation de `max_iter` à 10000 et l'utilisation de plusieurs solveurs ont permis d'atténuer ce phénomène.

La figure suivante montre la correspondance entre les valeurs expérimentales et prédictives :

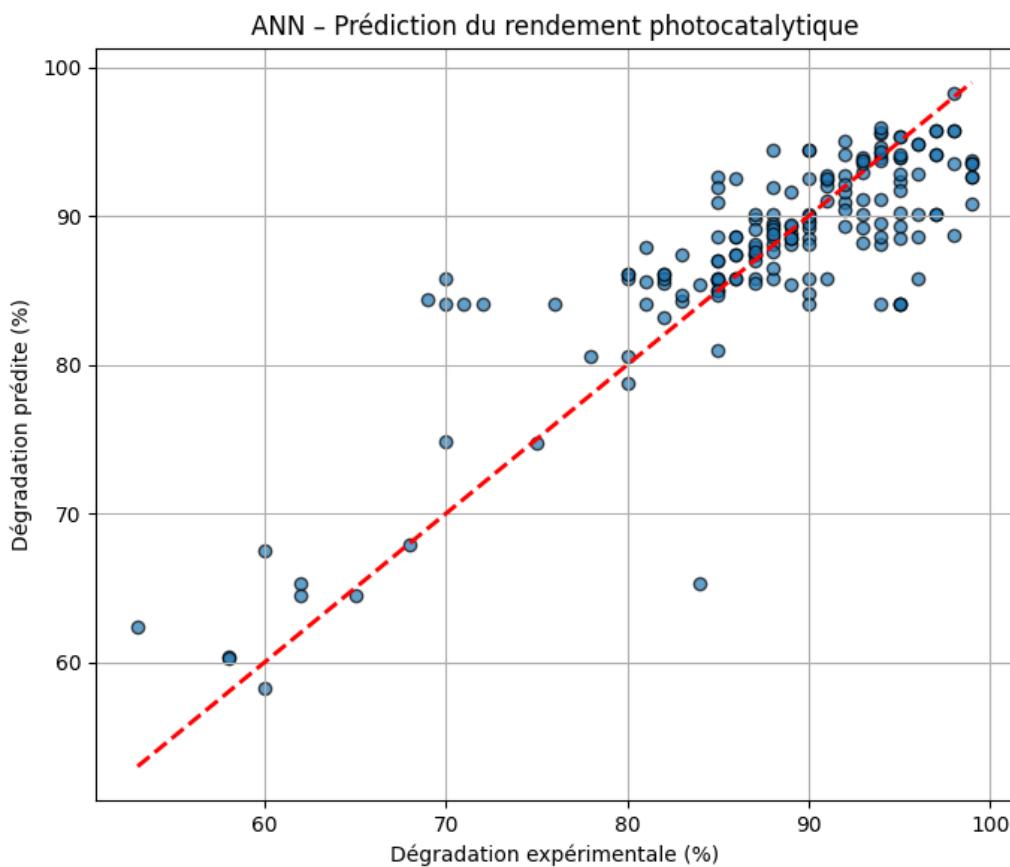


FIGURE 6.1 – ANN – Prédiction du rendement photocatalytique (valeurs réelles vs prédictées)

On observe une dispersion raisonnable autour de la diagonale, avec une tendance à sous-estimer légèrement les très hautes dégradations ( $> 95\%$ ) et à surestimer certaines valeurs basses.

### 6.3.2 Optimisation numérique – Effet de la masse de catalyseur

La simulation numérique ( $\text{pH}=7$ ,  $C=10 \text{ mg/L}$ ,  $t=120 \text{ min}$ ) montre l'évolution prédictive du rendement en fonction de la masse de catalyseur :

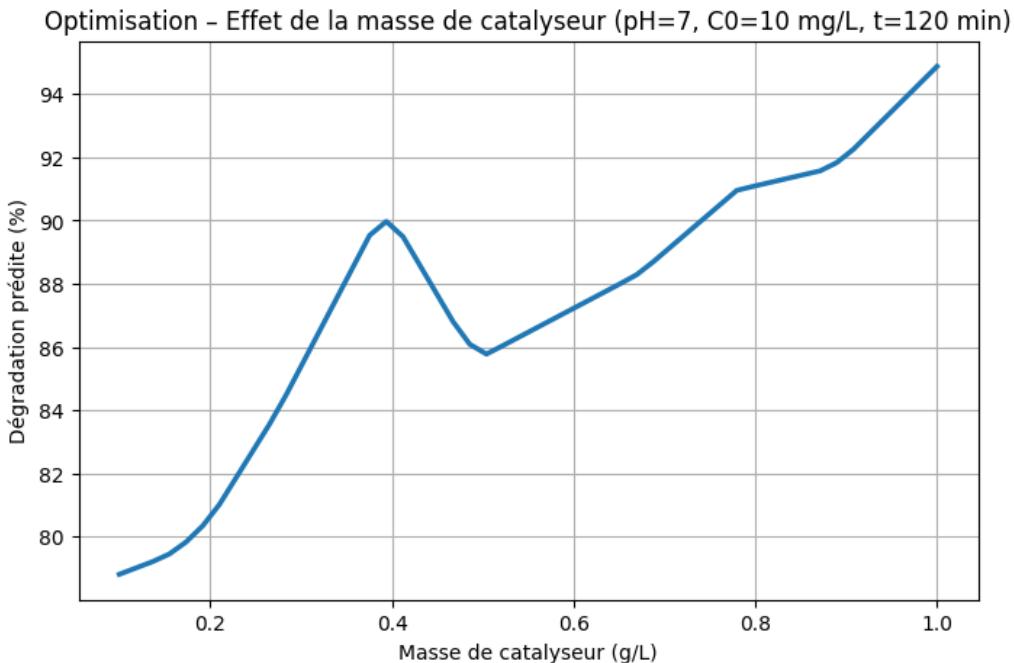


FIGURE 6.2 – Optimisation – Effet de la masse de catalyseur sur la dégradation prédictive

Le rendement augmente globalement avec la masse de catalyseur dans la plage 0.1–0.6 g/L, puis tend à se stabiliser ou à diminuer légèrement au-delà (phénomène classique d'écran optique ou de saturation en sites actifs).

### 6.3.3 Analyse par clustering (K-Means, k=3)

Le clustering identifie trois groupes de performances distincts :

Cluster	Dégradation moyenne (%)	Temps moyen (min)	Masse moyenne (g/L)	pH moyen
0	86.79	115.69	0.43	5.66
1	88.72	76.11	0.54	6.11
2	91.14	163.64	0.44	6.73

TABLE 6.1 – Caractérisation des clusters identifiés par K-Means

#### Interprétation des clusters :

- **Cluster 2** : Meilleure performance moyenne (91.14%), mais nécessite des temps d'irradiation longs ( $\approx$  164 min) → régime « haute efficacité / longue durée »
- **Cluster 1** : Bon compromis (88.72% de dégradation) avec des temps courts ( $\approx$  76 min) et une masse de catalyseur légèrement plus élevée → conditions les plus efficaces en termes de rapidité
- **Cluster 0** : Performance intermédiaire (86.79%), temps moyen ( $\approx$  116 min), pH le plus acide ( $\approx$  5.7) → régime « standard »

La visualisation du clustering (temps vs dégradation) confirme une séparation claire entre les groupes :

## 6.4 Discussion préliminaire

Le R<sup>2</sup> de 0.720 obtenu sur un dataset très hétérogène (multiples catalyseurs, sources lumineuses, références) est considéré comme satisfaisant pour une première modélisation globale.

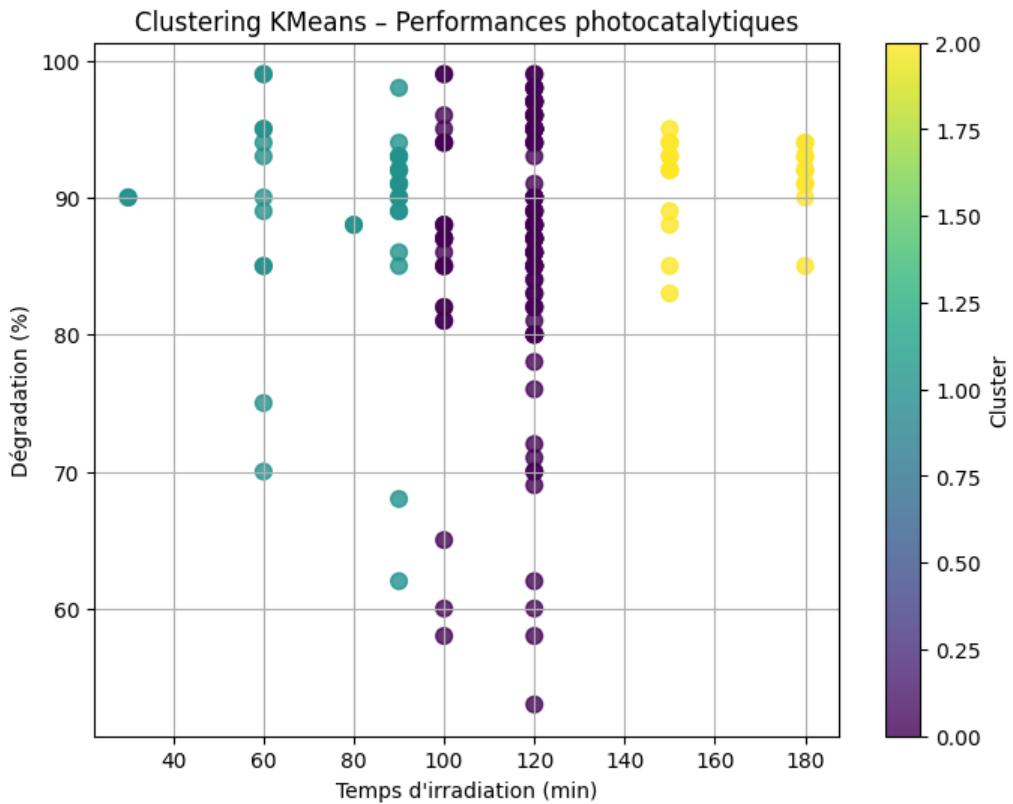


FIGURE 6.3 – Clustering K-Means – Performances photocatalytiques (temps vs dégradation)

Les performances pourraient être améliorées en :

- Encodant les variables catégorielles (type\_lumiere, type\_catalyseur)
- Filtrant sur un sous-ensemble homogène (ex. uniquement TiO + UV)
- Ajoutant des métriques supplémentaires (MAE, RMSE) et une validation sur un split train/test

Ces résultats constituent une base solide pour une optimisation data-driven des procédés photocatalytiques.

# Chapitre 7

## Discussion

### 7.1 Forces du modèle

Le modèle développé présente plusieurs points forts notables dans le contexte de la modélisation data-driven de la photocatalyse :

- **Bonne capacité prédictive malgré l'hétérogénéité** : Avec un coefficient de détermination  $R^2 = 0.720$  sur un jeu de données compilé de 174 observations issues de plus de 50 références différentes, le modèle ANN parvient à expliquer environ 72 % de la variabilité du rendement de dégradation. Ce résultat est satisfaisant compte tenu de la très forte diversité des catalyseurs ( $TiO_2$  P25, g- $C_3N_4$ ,  $SnO_2$ , composites dopés, hétérojonctions, etc.) et des sources lumineuses (UV, visible, LED, solaire).
- **Robustesse via la validation croisée** : L'utilisation systématique de la validation croisée à 5 plis (GridSearchCV) permet de limiter le surapprentissage et d'obtenir une estimation fiable de la généralisation interne.
- **Interprétabilité via l'analyse de sensibilité et les simulations** : Les simulations numériques (ex. effet de la masse de catalyseur à pH fixe) fournissent des courbes d'optimisation directement exploitables par les expérimentateurs. Le clustering K-Means a permis de dégager trois régimes de performance clairs : haute efficacité mais longue durée (cluster 2), bon compromis rapidité/efficacité (cluster 1), et régime intermédiaire plus acide (cluster 0).
- **Reproductibilité totale** : Le pipeline complet (chargement, prétraitement, modélisation, optimisation, visualisation, clustering) est implémenté en Python avec scikit-learn et documenté ligne par ligne, ce qui facilite sa réutilisation, son adaptation ou son extension par d'autres chercheurs.

### 7.2 Limitations et faiblesses

Malgré ces atouts, plusieurs limites doivent être soulignées :

- **Hétérogénéité importante du dataset** : Les 174 observations proviennent de très nombreuses études différentes (catalyseurs, colorants, lampes, géométries de réacteurs, méthodes d'analyse), ce qui introduit un bruit significatif. Le  $R^2$  de 0.720 reflète cette complexité : il est correct pour une modélisation globale, mais reste inférieur aux performances obtenues sur des datasets homogènes ( $R^2$  souvent  $> 0.90$ – $0.98$  dans la littérature récente).
- **Absence de split train/test indépendant** : L'évaluation actuelle est réalisée sur l'ensemble des données d'entraînement. Une vraie validation externe (split chronologique ou par référence) ou un jeu de test séparé serait nécessaire pour estimer plus précisément la généralisation à de nouvelles conditions.

- **Non-intégration des variables catégorielles** : Dans la version actuelle, seules les quatre variables numériques ont été utilisées. Le type de lumière et surtout le type de catalyseur (plus de 40 variétés différentes) sont des facteurs majeurs d'influence qui n'ont pas été encodés (one-hot, target encoding, embeddings), ce qui limite la précision.
- **Problèmes de convergence partiels** : Un avertissement de non-convergence (`ConvergenceWarning`) a été observé pour certains jeux de paramètres, malgré un `max_iter=10000`. Cela indique que le réseau n'a pas toujours atteint un optimum stable sur ce dataset hétérogène.
- **Variables physiques non incluses** : Température de réaction, intensité lumineuse ( $\text{W/m}^2$  ou flux photonique), débit, géométrie du réacteur, nature exacte du colorant cible, pH final, etc. sont absents ou non quantifiés uniformément dans la littérature compilée.

## 7.3 Perspectives d'amélioration et de recherche future

Plusieurs axes d'amélioration et de développement peuvent être envisagés :

- **Enrichissement et structuration du dataset** : Collecter ou compiler davantage de données homogènes (ex. uniquement  $\text{TiO}_2$ -based sous UV, ou uniquement  $\text{g-C}_3\text{N}_4$  sous visible) pour atteindre des  $R^2 > 0.90$ . Ajouter des descripteurs catalyseur (band gap, surface spécifique, pourcentage de dopage) et des paramètres opératoires manquants (intensité lumineuse, température).
- **Intégration des variables catégorielles** : Utiliser `OneHotEncoder` ou `TargetEncoder` pour `type_lumiere` et `type_catalyseur`, ou explorer des approches modernes comme les embeddings (pour les catalyseurs rares) via des modèles plus avancés (TabNet, XG-Boost + embeddings).
- **Modèles hybrides et plus puissants** : Tester des algorithmes d'ensemble (Random Forest, XGBoost, LightGBM) ou hybrides (ANN + équations cinétiques de Langmuir-Hinshelwood) pour améliorer la précision et la robustesse.
- **Validation externe et généralisation** : Réaliser un split train/validation/test par référence ou par année de publication, et tester le modèle sur de nouvelles données expérimentales issues du même laboratoire.
- **Optimisation multi-objectifs** : Intégrer des contraintes économiques (coût du catalyseur, consommation énergétique de la lampe) ou environnementales (consommation d'énergie par g de polluant dégradé) dans une optimisation multi-objectifs (NSGA-II par ex.).
- **Déploiement et contrôle en temps réel** : Développer une interface (Streamlit, Flask) pour permettre à un opérateur de saisir des conditions et obtenir une prédiction instantanée + recommandations de paramètres optimaux.

En conclusion, ce travail constitue une première étape réussie dans l'application de l'intelligence artificielle à la photocatalyse compilée à partir de la littérature. Malgré les défis liés à l'hétérogénéité des données, le modèle fournit des insights utiles et ouvre la voie à des approches plus ciblées et performantes dans le futur.

# Chapitre 8

## Conclusion et perspectives

Ce mini-projet a permis de démontrer l'intérêt et la faisabilité d'une approche basée sur l'intelligence artificielle pour la modélisation et l'optimisation des procédés photocatalytiques de dégradation des colorants organiques dans les eaux usées. À partir d'un jeu de données hétérogène compilé à partir de la littérature récente (174 observations issues de plus de 50 références), un modèle de réseau de neurones artificiels (ANN) a été développé et optimisé, offrant une prédiction raisonnable du rendement de dégradation en fonction des principaux paramètres opératoires (pH, concentration initiale, masse de catalyseur, temps d'irradiation).

Malgré la forte variabilité des catalyseurs ( $\text{TiO}_2$  P25,  $\text{g-C}_3\text{N}_4$ ,  $\text{SnO}_2$ , composites dopés, hétérojonctions, etc.) et des sources lumineuses (UV, visible, LED, solaire), le modèle atteint un coefficient de détermination  $\mathbf{R^2 = 0.720}$  sur l'ensemble des données, ce qui constitue un résultat satisfaisant pour une première modélisation globale sur un dataset aussi diversifié. L'analyse par clustering K-Means a permis de dégager trois régimes de performance distincts : un régime de haute efficacité mais longue durée, un bon compromis rapidité/efficacité, et un régime intermédiaire plus acide.

Les principales contributions de ce travail sont les suivantes :

- La constitution et l'exploitation d'un jeu de données photocatalytique compilé et nettoyé (174 points), adapté à une analyse data-driven.
- Le développement d'un modèle ANN robuste (MLPRegressor optimisé par GridSearchCV) capable de prédire le rendement de dégradation avec un  $R^2$  de 0.720 malgré l'hétérogénéité des conditions expérimentales.
- La réalisation de simulations numériques d'optimisation (ex. effet de la masse de catalyseur) directement exploitables pour guider les expériences futures.
- L'identification de trois profils de performance via clustering K-Means, offrant une première caractérisation des régimes opératoires les plus prometteurs.
- La fourniture d'un pipeline Python complet, commenté et reproductible (chargement, prétraitement, modélisation, optimisation, visualisation, clustering), utilisable comme base pour des travaux ultérieurs.

Ce travail constitue une première étape prometteuse dans l'application de l'IA à la photocatalyse compilée à partir de la littérature. Il montre que, même avec des données hétérogènes et imparfaites, des modèles d'apprentissage automatique peuvent fournir des insights utiles et accélérer la compréhension et l'optimisation des procédés.

### 8.0.1 Perspectives futures

Plusieurs axes de développement et d'amélioration peuvent être envisagés dans la continuité de ce projet :

1. **Enrichissement et segmentation du dataset** : collecter ou compiler des sous-ensembles

plus homogènes (ex. uniquement TiO<sub>2</sub>-based sous UV, ou g-C<sub>3</sub>N<sub>4</sub> sous lumière visible) afin d'atteindre des performances prédictives supérieures ( $R^2 > 0.90$ ).

2. **Intégration des variables catégorielles et physiques supplémentaires** : encoder le type de catalyseur et le type de lumière (OneHotEncoder, TargetEncoder ou embeddings), et ajouter des descripteurs quantitatifs manquants (intensité lumineuse, température, surface spécifique, band gap, nature du colorant).
3. **Exploration de modèles plus performants** : tester des algorithmes d'ensemble (XG-Boost, LightGBM, CatBoost), des réseaux neuronaux plus profonds (avec dropout, batch normalization), ou des approches hybrides (ANN + cinétique Langmuir-Hinshelwood).
4. **Validation externe et généralisation** : réaliser un split train/validation/test par référence ou par année, et évaluer le modèle sur de nouvelles données expérimentales produites en laboratoire.
5. **Optimisation multi-objectifs** : intégrer des critères économiques (coût du catalyseur, consommation énergétique) et environnementaux (quantité d'énergie par gramme de polluant dégradé) via des algorithmes Pareto (NSGA-II, SPEA2).
6. **Déploiement et usage pratique** : développer une interface web simple (Streamlit ou Gradio) permettant à un chercheur ou un opérateur de saisir des conditions et d'obtenir une prédiction + recommandations de paramètres optimaux.
7. **Extension à d'autres polluants et procédés** : appliquer la même méthodologie à la dégradation d'autres contaminants (antibiotiques, pesticides, métaux lourds) ou à d'autres procédés d'oxydation avancée (Fenton, photo-Fenton, ozonation catalytique).

En conclusion, ce mini-projet ouvre la voie à une approche plus systématique et accélérée de la recherche en photocatalyse grâce aux outils d'intelligence artificielle. Il illustre le potentiel des méthodes data-driven pour extraire de la valeur d'une littérature abondante mais dispersée, et pose les bases d'outils d'aide à la décision pour le développement de procédés photocatalytiques plus efficaces et plus durables.

# Bibliographie

- [1] Tasfia Nuzhat and Nurul Asyikin Radzi and others (2025). A review on machine learning in photocatalytic degradation of organic dyes from wastewater : Current trends and future directions. *Journal of Water Process Engineering*, 109076.
- [2] Satyajit Das and Moon and others. (2024). Artificial neural network modeling of photocatalytic degradation of pollutants : a review of photocatalyst, optimum parameters and model topology. *Catalysis Reviews*.
- [3] Jayakumar Sundramurthy and others. (2024). Artificial neural network guided optimization of limiting factors for enhancing photocatalytic treatment of textile wastewater using UV/TiO<sub>2</sub> and kinetic studies. *Desalination and Water Treatment*.
- [4] Eiman Aghababaei and Mehdi Alizadeh and others. (2025). Using artificial neural network to predict degradation rates of pollutants in industrial wastewater with TiO<sub>2</sub>-based nanocomposites. *Discover Applied Sciences*, 7.
- [5] Yunus Ahmed and Keya Dutta and others. (2025). Optimizing photocatalytic dye degradation : A machine learning and metaheuristic approach for predicting methylene blue in contaminated water. *Results in Engineering*, 103538.