الجمهوريه الجزائريه الديموقراطيه الشعبيه

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

N° Réf:……

Université Abdelhafid Boussouf - Mila
Institut de Mathématiques et Informatique

Département d'Informatique

Mémoire de Fin d'Etude En vue de l'obtention du diplôme de Master

En : **Informatique**

Spécialité : Intelligence Artificielle et ses Applications (I2A)

Thème

Prédiction du Crop Yield à haute précision par des modèles Machine Learning et Deep Learning pour une agriculture intelligente et durable

Préparé par :

- **BOUHENNACHE Chaima**
- **DEFFOUS Sihem**

Examiné par le jury :

Dr. GUEMRI Oualid	**MCA**	**Président**
Dr. BEKHOUCHE Maamar	**MCB**	**Examinateur**
Dr. HADJI Atmane	**MCB**	**Superviseur**

Année Universitaire : 2025/2026

Remerciements

Tout d’abord, nous tenons à remercier Allah le Tout-Puissant de nous avoir accordé la force, la patience et le courage nécessaires pour mener à bien ce modeste travail.

Nous exprimons notre profonde gratitude à notre encadrant, **Dr. HADJI Atmane**, pour avoir accepté de diriger ce mémoire. Ses précieux conseils, sa rigueur scientifique et sa disponibilité constante ont été pour nous une source d'apprentissage inestimable.

Nos remerciements s'adressent également aux membres du jury qui ont accepté d'évaluer notre travail et de nous faire l'honneur de présider et d'examiner cette soutenance.

À nos chères familles, en particulier à nos parents, pour leur amour inconditionnel, leurs sacrifices et leurs prières qui nous ont portées tout au long de notre cursus universitaire. Ce travail est le fruit de votre soutien indéfectible.

Enfin, un merci tout particulier à ma binôme et amie de route. Ce mémoire n'est pas seulement un projet académique, c'est le témoignage d'une collaboration exemplaire, de moments de partage, de doutes surmontés ensemble et d'une amitié renforcée par l'effort. Merci pour ta confiance et pour cette solidarité qui a rendu ce voyage plus beau.

À tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce projet, merci.

BOUHENNACHE Chaima & DEFFOUS Sihem

Liste des sigles et acronymes

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
LSTM	Long Short
XGBoost	eXtreme
LightGBM	Light Gradient BoostingMachine
RF	Random Forest
SVR	Support Vector Regression
CatBoost	Categorical Boosting
MLP	Multi Layer Perceptron
GRU	GatedRecurrent Unit
NDVI	Normalized Difference Vegetation Index
EVI	Enhanced Vegetation Index
LAI	Leaf Area Index
FPAR	Fraction of Photosynthetically Active Radiation
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
R^2	Coefficient of Determination
EDA	Exploratory Data Analysis
SHAP	SHapley Additive Explanation

Résumé

La prévision du rendement agricole constitue un enjeu majeur dans le contexte de l'agriculture intelligente, où l'optimisation des ressources et l'amélioration de la productivité sont essentielles face aux défis climatiques et environnementaux. Dans ce cadre, l'intelligence artificielle, notamment le Machine Learning et le Deep Learning, offre des solutions performantes pour modéliser les relations complexes entre les facteurs influençant la production agricole.

Ce mémoire a pour objectif de concevoir, d'implémenter et de comparer différents modèles intelligents afin de prédire le rendement agricole à partir du jeu de données **CropYield**, composé de **90 336 observations**. Ce jeu de données intègre des données multi-sources, notamment des variables climatiques (température et précipitations), des indices de végétation issus de la télédétection (NDVI, EVI, LAI et FPAR), des données radar (VV, VH) ainsi que des informations relatives à l'humidité du sol et à l'évaporation.

Plusieurs approches ont été étudiées, allant des modèles classiques tels que SVR et Random Forest aux méthodes de boosting (XGBoost, LightGBM, Gradient Boosting et CatBoost), ainsi qu'aux modèles de Deep Learning (MLP, LSTM et GRU).

Afin d'améliorer davantage les performances prédictives, une contribution de ce travail consiste en la proposition et le développement d'un modèle hybride innovant combinant LightGBM, CatBoost et un réseau de neurones LSTM. Cette architecture permet d'exploiter simultanément les capacités des modèles de boosting pour la modélisation des relations non linéaires et celles du LSTM pour l'extraction de représentations pertinentes des données, améliorant ainsi la précision et la robustesse des prédictions.

Un processus complet de prétraitement des données a été réalisé, comprenant le traitement des valeurs manquantes, l'encodage des variables catégorielles, l'intégration des données hétérogènes, ainsi que l'ingénierie des caractéristiques. Les modèles ont été évalués à l'aide des métriques MSE, RMSE, MAE, MAPE, MedAE, R^2, EVS et Adjusted R^2.

Les résultats expérimentaux montrent que le modèle hybride proposé surpasse l'ensemble des modèles comparés en obtenant les meilleures performances, avec notamment un coefficient de détermination **R^2 = 0,82874**, démontrant ainsi sa capacité à expliquer près de **82,9 %** de la variabilité du rendement agricole. Ces résultats confirment l'intérêt de combiner les approches de Machine Learning et de Deep Learning au sein d'une architecture hybride afin d'améliorer la précision, la robustesse et la capacité de généralisation des systèmes de prévision du rendement agricole.

Mots-clés : Agriculture intelligente, Prévision du rendement agricole, Intelligence artificielle, Machine Learning, Deep Learning, Modèle hybride, LightGBM, CatBoost, LSTM.

Abstract

Crop yield prediction has become a major challenge in the context of smart agriculture, where resource optimization and productivity improvement are essential to address climate and environmental challenges. In this context, Artificial Intelligence (AI), particularly Machine Learning and Deep Learning, provides powerful solutions for modeling the complex relationships among the factors influencing agricultural production.

This thesis aims to design, implement, and compare different intelligent models for crop yield prediction using the **CropYield** dataset, which consists of **90,336 observations**. The dataset integrates multi-source information, including climatic variables (temperature and rainfall), vegetation indices derived from remote sensing (NDVI, EVI, LAI, and FPAR), radar data (VV and VH), as well as soil moisture and evaporation variables.

Several approaches were investigated, ranging from traditional Machine Learning models such as SVR and Random Forest to boosting algorithms including XGBoost, LightGBM, Gradient Boosting, and CatBoost, as well as Deep Learning models such as MLP, LSTM, and GRU.

To further improve predictive performance, this work proposes and develops an innovative hybrid model that combines **LightGBM**, **CatBoost**, and an **LSTM** neural network. This hybrid architecture leverages the strengths of boosting algorithms in modeling complex nonlinear relationships while benefiting from the LSTM network's ability to learn meaningful representations from the data, thereby improving prediction accuracy, robustness, and generalization capability.

A comprehensive data preprocessing pipeline was performed, including missing value handling, categorical variable encoding, heterogeneous data integration, and feature engineering. The developed models were evaluated using several performance metrics, including MSE, RMSE, MAE, MAPE, MedAE, R^2, EVS, and Adjusted R^2.

The experimental results demonstrate that the proposed hybrid model outperformed all the compared models, achieving the best predictive performance with a **coefficient of determination (R^2) of 0.82874**, indicating that it explains approximately **82.9%** of the variability in crop yield. These findings highlight the effectiveness of integrating Machine Learning and Deep Learning techniques within a hybrid architecture to enhance the accuracy, robustness, and generalization of crop yield prediction systems.

Keywords: Smart Agriculture, Crop Yield Prediction, Artificial Intelligence, Machine Learning, Deep Learning, Hybrid Model, LightGBM, CatBoost, LSTM.

الملخص

يُعد التنبؤ بالمردود الزراعي من أبرز التحديات في مجال الزراعة الذكية، حيث يُسهم في تحسين استغلال الموارد وزيادة الإنتاجية لمواجهة التحديات المناخية والبيئية. وفي هذا الإطار، يوفر الذكاء الاصطناعي، ولا سيما التعلم الآلي (Machine Learning) والتعلم العميق(Deep Learning) ، حلولًا فعّالة لنمذجة العلاقات المعقدة بين العوامل المؤثرة في الإنتاج الزراعي.
يهدف هذا العمل إلى تصميم وتنفيذ ومقارنة مجموعة من النماذج الذكية للتنبؤ بالمردود الزراعي اعتمادًا على مجموعة بيانات **CropYield**التي تضم **90,336 عينة** .وتجمع هذه المجموعة بيانات متعددة المصادر تشمل المتغيرات المناخية (درجة الحرارة وتساقط الأمطار)، ومؤشرات الغطاء النباتي المستخرجة من الاستشعار عن بعد (NDVI و EVIو LAIوFPAR)، وبيانات الرادار(VV وVH) الإضافة إلى بيانات رطوبة التربة والتبخر.
تمت دراسة عدة نماذج، بدءًا من النماذج التقليدية مثل SVR وRandom Forest، مرورًا بخوارزميات التعزيز (Boosting) مثل XGBoost و LightGBM و Gradient Boosting وCatBoost، وصولًا إلى نماذج التعلم العميق مثل MLP و LSTMوGRU.
ولتحسين دقة التنبؤ، تم اقتراح وتطوير نموذج هجين مبتكر يجمع بين **LightGBM** و **CatBoost** وشبكة **LSTM** العصبية. ويستفيد هذا النموذج من قدرة خوارزميات التعزيز على نمذجة العلاقات غير الخطية، ومن قدرة شبكة LSTM على تعلم واستخراج تمثيلات ذات معنى من البيانات، مما يسهم في تحسين دقة التنبؤ، وزيادة متانة النموذج، وتعزيز قدرته على التعميم.
كما تم تنفيذ مرحلة شاملة لمعالجة البيانات، شملت معالجة القيم المفقودة، وترميز المتغيرات الفئوية، ودمج البيانات متعددة المصادر، وهندسة الخصائص. وتم تقييم النماذج باستخدام مجموعة من مؤشرات الأداء وهي MSE وRMSE و MAE و MAPEو MedAE و R^2و EVS وR^2 Adjusted.
أظهرت النتائج التجريبية أن النموذج الهجين المقترح تفوق على جميع النماذج الأخرى، محققًا أفضل أداء، حيث بلغ **معامل التحديد (R^2) قيمة 0.82874**، مما يدل على قدرته على تفسير حوالي **82.9**٪ من التباين في المردود الزراعي. وتؤكد هذه النتائج أهمية دمج تقنيات التعلم الآلي والتعلم العميق ضمن بنية هجينة لتحسين دقة التنبؤ، وزيادة متانة النماذج، وتعزيز قدرتها على التعميم في مختلف البيئات الزراعية.

الكلمات المفتاحية :الزراعة الذكية، التنبؤ بالمردود الزراعي، الذكاء الاصطناعي، التعلم الآلي، التعلم العميق، النموذج الهجين، LightGBM، CatBoost، LSTM.

Table des matières

Remerciements I
Liste des sigles et acronymes II
Résumé III
Abstract IV
Liste des Figures X
Liste des Tableaux XI

Introduction générale 1

Chapitre 1 : System de prédiction pour l'Agriculture Intelligente

1.1 Introduction 5
1.2 Agriculture Intelligente 5
1.3 Importance de la Prédiction Agricole 5
1.4 Sources de Données Agricoles 6
1.5 Système de Prédiction Agricole 6
1.6 Applications des Systèmes de Prédiction Agricole 7
1.7 Défis et Perspectives 8
1.8 Conclusion 8

Chapitre 2 : Intelligence Artificielle et Télédétection pour l'Agriculture Intelligente

2.1 Introduction 10
2.2 Contexte et Enjeux des Céréales en Algérie 10
2.2.1 Agriculture Intelligente et Prédiction du Rendement 10
2.2.2 Les Céréales en Algérie : Enjeux Stratégiques 11
2.3 Télédétection : Source de Données Fondamentale pour les Céréales 11
2.3.1 Principes et Principaux Satellites Utilisés 11
2.3.2 Données Radar (Sentinel-1) et Fusion Multi-Sources 12

2.4 Indices de Végétation pour le Suivi des Céréales....13
2.5 Modèles de Machine Learning pour la Prédiction du Rendement Céréalier....13
2.5.1 Vue d'Ensemble....13
2.5.2 Random Forest....13
2.5.3 Modèles de Boosting : XGBoost et LightGBM....14
2.5.4 Autres Modèles : SVR et CatBoost....14
2.5.5 Limites des Modèles de Machine Learning....14
2.6 Deep Learning pour la Prédiction du Rendement Céréalier....15
2.6.1 Architecture LSTM....15
2.6.2 Architecture *GRU*....*15*
2.6.3 Architectures Hybrides et Avancées....15
2.7 Comparaison ML vs Deep Learning pour les Céréales....16
2.8 Intelligence Artificielle Explicable (XAI) pour les Céréales....16
2.8.1 Tableau Comparatif des Travaux Récents....17
2.9 Synthèse, Lacunes et Apport du Travail....17
2. 10 Conclusion....18

Chapitre 3 : Approche hybride proposée (LSTM-LightGBM-CatBoost)

3.1 Introduction....20
3.2 Méthodologie de l'approche proposée (hybride LSTM+ LightGBM+CatBoost)....20
3.3 Choix du Dataset....21
3.3.1 Données climatiques et du sol (Climate & Soil Data)....21
3.3.2 Données radar (Radar Data – Sentinel-1)....22
3.3.3 Indices de végétation (Vegetation Indices – Sentinel-2)....22
3.3.4 Données de production et variables catégorielles....22
3.4 Présentation du dataset....23
3.5 Visualisation de dataset....24
3.5.1 Partitionnement des données....24
3.5.3 Analyse statistique et corrélation....27
3.6 Prétraitement des données....27
3.7 Ingénierie des fonctionnalités....30
3.7.1 Extraction des indices de végétation et des données radar....30

3.7.2 Encodage des variables catégorielles 30
3.7.3 Encodage des variables temporelles 30
3.7.4 Intégration des données provenant de plusieurs sources (Fusion de don- nées) 31
3.8 Sélection des variables 31
3.9 Paramétrage des Algorithmes 32
3.9.1 Hyperparamètres spécifiques pour chaque modèle *32*
3.9.2 Validation et réglage des modèles *32*
3.9.3 Gestion de l'apprentissage et optimisation des modèles 33
3.9.4 Métriques d'évaluation 33
3.10 Description des modèles Machine learning 34
3.10.1 Modèles d'apprentissage supervisé 35
3.10.2 Algorithmes de Gradient Boosting 35
3.11 Description des modèles deep learning 38
3.12 Techniques avancées 41
3.13 Concluction 44

Chapitre 4 : Implémentation et expérimentation

4.1 Introduction 46
4.2 Environnement de développement 46
4.2.1 Langage de programmation 46
4.2.2 Bibliothèques de Machine Learning 46
4.2.3 Bibliothèques de Deep Learning 46
4.2.4 Traitement et visualisation des données 47
4.2.5 Outils complémentaires 47
4.3 Évaluation des performances 47
4.3.1 Résultats des modèles ML 48
4.3.2 Résultats des modèles DL 51
4.3.3 Résultat de modèle hybride 53
4.4 Comparaison des performances 54
4.4.1 Comparaison des performances selon les algorithmes 54
Shape 55
4.5 Application Web 56

4.5.1 Présentation Générale de l'Application 56
4.5.2 Développement de l'Application 56
4.6 Développement du Modèle Final 57
4.6.1 Sérialisation et Persistance avec Joblib 57
4.6.2 Pipeline de Prétraitement et Encodage des Variables 58
4.7 Développement de l'API Backend 59
4.7.1 Justification du Choix de FastAPI 59
4.7.2 Cycle de Traitement du Point de Terminaison /predict 59
4.7.3 Sécurité et Documentation (Swagger UI) 60
4.8 Développement de l'Interface Utilisateur (Frontend) 60
4.8. 1 Streamlit et Choix Technologiques 60
4.8.2 Organisation Ergonomique de l'Interface 61
4.8.3 Visualisations Avancées et Exportation 61
4.8.4 Présentation des Interfaces et Fonctionnalités 62
4.10 Conclusion 67
Conclusion générale 69
Défis du travail 70
Perspectives 71
Bibliographie 72

Liste des Figures

Figure 1: Système de Prédiction Agricole 7
Figure 2: Méthodologie globale de notre système 21
Figure 3: Capture d'écran de Dataset 24
Figure 4: La distribution d'observations par wilaya 25
Figure 5 : Analyse Spatiale de rendement par wilaya 25
Figure 6 : Évolution du rendement par wilaya 26
Figure 7 : Matrice de corrélation 27
Figure 8 : Architecture du modèle RF 35
Figure 9 : Architecture du modèle SVR 35
Figure 10 : Architecture du modèle XGBoost 36
Figure 11 : Architecture du modèle LightGBM 37
Figure 12 : Architecture du modèle CatBoot 37
Figure 13 : Architecture du modèle LSTM 39
Figure 14 : Architecture du modèle GRU 40
Figure 15 : Architecture du modèle MLP 41
Figure 16 : Architecture du modèle hybride 43
Figure 17 : Évaluation Random Forest 48
Figure 18 : Évaluation LightGBM 49
Figure 19 : Évaluation XGBoost 49
Figure 20 : Évaluation GBR 50
Figure 21: Évaluation CatBoost 50
Figure 22 : Évaluation SVR 51
Figure 23 : Évaluation LSTM 52
Figure 24 : Évaluation MLP 52
Figure 25 : Évaluation GRU 53
Figure 26 : Évaluation de notre modèle proposé hybride 54
Figure 27: SHAP Beeswarm et Importance globale 56
Figure 28 : Diagramme de flux détaillé 62
Figure 29 : L'interface d'accueil AgriPredict 63
Figure 30 : Page À propos du Modèle 63
Figure 31 : Interface de saisie des données et système d'aide interactive (Infobulles). 64
Figure 32 : Corrélation et alertes intelligentes basées sur les spécificités climatiques. 65
Figure 33 : Courbe de croissance prévisionnelle basée sur le modèle sigmoïde (Octobre - Juin). 65
Figure 34 : Analyse de l'importance des variables via les valeurs SHAP. 66
Figure 35 : Certificat technique de prévision du rendement généré au format PDF 67

Liste des Tableaux

Tableau 1: Comparaison des principales études sur la prédiction du rendement agricole par IA et Télédétection 17
Tableau 2: Description des attributs du dataset 23
Tableau 3 : Résultats de performance des modèles de Machine Learning 48
Tableau 4: Résultas de performance des modèles de Deep Learning 51
Tableau 5: Résultat de performance de modèle hybride 53
Tableau 6: Evaluation des meilleurs modèles ML et hybride 54

Introduction générale

L'agriculture constitue un pilier fondamental de l'économie mondiale et un élément essentiel pour assurer la sécurité alimentaire des populations. Cependant, ce secteur stratégique est aujourd'hui confronté à de nombreux défis, notamment les changements climatiques, la rareté des ressources en eau, ainsi que la dégradation progressive des sols. Ces facteurs influencent directement la productivité agricole et rendent la gestion des cultures de plus en plus complexe. Dans ce contexte, la prévision précise du rendement agricole devient une nécessité stratégique, permettant d'optimiser les pratiques agricoles, de réduire les risques et d'améliorer la prise de décision.

Avec l'émergence de l'agriculture intelligente, basée sur l'intégration des technologies numériques avancées, les méthodes agricoles traditionnelles ont connu une transformation significative. Cette nouvelle approche repose sur l'exploitation de données massives provenant de sources variées telles que les données climatiques, satellitaires et spatiales. L'intelligence artificielle, en particulier, joue un rôle central dans cette évolution, en offrant des outils capables d'analyser ces données complexes et d'en extraire des informations pertinentes pour améliorer la productivité agricole.

L'agriculture intelligente vise ainsi à optimiser l'utilisation des ressources (eau, énergie, engrais) tout en augmentant la qualité et la quantité de la production. Elle s'appuie sur des technologies telles que les capteurs, les systèmes de positionnement (GPS), la télédétection, les technologies de l'information et de la communication (TIC), ainsi que l'intelligence artificielle, afin de surveiller les cultures en temps quasi réel et d'assister les agriculteurs dans leurs décisions stratégiques et opérationnelles.

Dans ce cadre, l'estimation du rendement agricole représente un enjeu majeur. En effet, le rendement dépend de nombreux facteurs interdépendants, tels que les conditions météorologiques, les caractéristiques du sol, les pratiques agricoles et les indices issus de la télédétection, comme le NDVI ou l'EVI. Cette complexité rend les approches classiques de prévision, souvent basées sur des modèles statistiques ou empiriques, limitées en termes de précision et d'adaptabilité. De plus, ces méthodes ne permettent pas toujours de modéliser efficacement les relations non linéaires entre les variables ni d'exploiter pleinement les données issues de sources multiples.

Face à ces limitations, les techniques d'intelligence artificielle, notamment le Machine Learning et le Deep Learning, offrent des solutions prometteuses. Ces approches permettent de modéliser des relations complexes, d'intégrer de grandes quantités de données hétérogènes et d'améliorer significativement la précision des prédictions. Dans ce contexte, la problématique

centrale de ce travail consiste à explorer comment exploiter ces techniques afin de prédire avec précision le rendement agricole en se basant sur des données climatiques, satellitaires et spatiales.

Dans le cadre de ce travail, une contribution a été apportée à travers le développement d'un modèle hybride proposé spécifiquement dans cette étude. Contrairement aux approches classiques qui reposent sur un seul algorithme, ce modèle combine les capacités des réseaux LSTM à capturer les relations temporelles présentes dans les données avec les performances des algorithmes de boosting LightGBM et CatBoost, afin d'améliorer davantage la précision et la robustesse des prédictions du rendement agricole.

L'objectif principal de ce projet est donc de développer un modèle performant capable d'estimer le rendement agricole de manière fiable. Pour atteindre cet objectif, plusieurs étapes sont nécessaires, notamment l'analyse et le prétraitement des données, la sélection des variables pertinentes, le développement de modèles de Machine Learning (tels que Random Forest ou XGBoost), de Deep Learning (notamment les réseaux LSTM), ainsi que de modèles hybrides, et enfin l'évaluation et la comparaison de leurs performances à l'aide de métriques appropriées telles que RMSE, MAE et R^2.

Au-delà de l'aspect prédictif, ce projet vise également à concevoir un système intelligent d'aide à la décision, permettant de valoriser les résultats obtenus et de les rendre exploitables par les utilisateurs finaux, tels que les agriculteurs ou les décideurs. Ce système contribuera à améliorer la gestion des cultures et à soutenir une agriculture plus durable et plus efficace.

Par ailleurs, ce travail apporte plusieurs contributions. D'un point de vue scientifique, il propose une approche intégrée combinant différentes sources de données (climatiques, satellitaires et spatiales) afin d'obtenir une vision globale des facteurs influençant le rendement agricole. Il met également en œuvre des modèles avancés capables de capturer les relations complexes entre ces variables. L'introduction de modèles hybrides constitue également une contribution majeure, permettant de combiner les forces du Machine Learning et du Deep Learning. De plus, une analyse comparative des performances des différents modèles est réalisée afin d'identifier les approches les plus efficaces. L'intégration de la dimension temporelle, à travers l'utilisation de modèles comme les réseaux LSTM, constitue également un apport significatif de ce projet.

D'un point de vue pratique, ce travail se distingue par le développement d'une solution complète, allant de l'analyse des données à la mise en œuvre d'un système intelligent. Cette approche permet non seulement de prédire le rendement agricole, mais aussi d'accompagner les utilisateurs dans leur prise de décision. Elle ouvre également des perspectives d'amélioration futures, notamment par l'intégration de techniques d'intelligence artificielle explicable (XAI) et

l'optimisation des performances des modèles.

Afin de répondre aux objectifs de cette étude, ce mémoire est structuré en quatre chapitres complémentaires, organisés de manière progressive pour couvrir l'ensemble des aspects liés à la prévision du rendement agricole dans le contexte de l'agriculture intelligente.

Chapitre 1 : Présente le contexte de l'agriculture intelligente ainsi que les enjeux liés à la prédiction du rendement. Il introduit également le rôle de l'intelligence artificielle dans l'amélioration des systèmes agricoles.

Chapitre 2 : Synthétise les travaux existants sur la prédiction du rendement agricole. Il met en évidence les principales approches proposées dans la littérature ainsi que leurs limites.

Chapitre 3: Décrit la méthodologie adoptée et les données utilisées. Il présente également les modèles d'intelligence artificielle développés dans le cadre de cette étude.

Chapitre 4 : Présente l'implémentation du système proposé ainsi que les résultats expérimentaux obtenus. Il analyse et compare les performances des différents modèles.

Chapitre 1

System de prédiction pour l'Agriculture Intelligente

1.1 Introduction

L'agriculture occupe une place essentielle dans le développement économique et social des pays. Elle constitue la principale source de production alimentaire et contribue à la sécurité alimentaire mondiale [1]. Cependant, ce secteur est aujourd'hui confronté à de nombreux défis tels que la croissance démographique, les changements climatiques, la diminution des ressources naturelles et l'augmentation de la demande en produits agricoles [2]. Ces contraintes rendent nécessaire l'adoption de nouvelles approches capables d'améliorer la gestion des exploitations agricoles et d'optimiser la production.

Dans ce contexte, l'agriculture intelligente apparaît comme une solution innovante reposant sur l'utilisation des technologies numériques et des systèmes d'aide à la décision [3]. Grâce à l'exploitation de données provenant de différentes sources, il devient possible de suivre l'évolution des cultures, d'anticiper les risques et d'améliorer la productivité agricole. Parmi les applications les plus importantes de cette approche figure la prédiction du rendement agricole [4].Ce chapitre présente les concepts fondamentaux liés à l'agriculture intelligente ainsi qu'aux systèmes de prédiction agricole.

1.2 Agriculture Intelligente

L'agriculture intelligente représente une évolution des pratiques agricoles traditionnelles basée sur l'intégration des technologies de l'information et de la communication (TIC) dans les activités agricoles [3]. Elle vise à améliorer la productivité tout en assurant une gestion durable des ressources naturelles.

Cette approche repose sur la collecte et l'analyse de données provenant de différentes sources afin d'optimiser les décisions agricoles. L'utilisation des technologies numériques permet de mieux surveiller les parcelles et de réagir rapidement aux conditions environnementales changeantes.

L'agriculture intelligente est fortement liée à l'Internet des objets (IoT), au Big Data et à l'intelligence artificielle [5]. Ces technologies permettent de transformer les données agricoles en informations exploitables.

1.3 Importance de la Prédiction Agricole

La prédiction agricole consiste à estimer à l'avance différents paramètres liés à la production agricole, notamment le rendement des cultures. Elle constitue un outil stratégique permettant d'améliorer la gestion des ressources agricoles et de réduire les incertitudes associées aux activités de production [6].

L'importance de la prédiction agricole s'est considérablement accrue avec l'augmentation des

risques liés aux changements climatiques. Les variations de température, les sécheresses et les irrégularités des précipitations influencent directement la croissance des cultures et rendent les rendements plus difficiles à prévoir. Dans ce contexte, les modèles de prédiction permettent de fournir des estimations précoces susceptibles d'améliorer la planification agricole et la gestion des exploitations [2].

Enoutre, les prévisions de rendement aident les agriculteurs à optimiser l'utilisation des ressources telles que l'eau, les engrais et les produits phytosanitaires. Elles contribuent également à réduire les coûts de production et à améliorer l'efficacité des pratiques agricoles [3].

Les résultats issus des systèmes de prédiction peuvent aussi être exploités par les organismes publics pour anticiper les niveaux de production, planifier les besoins alimentaires et soutenir les politiques agricoles. Ainsi, la prédiction agricole joue un rôle important dans l'amélioration de la productivité et le renforcement de la sécurité alimentaire [1].

1.4 Sources de Données Agricoles

La qualité et la fiabilité des prédictions agricoles dépendent fortement des données utilisées pour décrire l'environnement de production et l'état des cultures.

Les données climatiques constituent l'une des principales sources d'information. Elles regroupent différents paramètres météorologiques tels que la température, les précipitations, l'humidité de l'air et le rayonnement solaire, qui influencent directement la croissance et le développement des cultures [2].

Les données satellitaires jouent également un rôle important dans le suivi des parcelles agricoles. Grâce aux images acquises par les satellites, il est possible d'observer l'état de la végétation et l'évolution des cultures [7].

Les données agronomiques fournissent des informations relatives aux pratiques agricoles et aux rendements enregistrés.

Enfin, **les caractéristiques du sol** influencent la disponibilité de l'eau et des nutriments nécessaires à la croissance des plantes [1].

L'intégration de ces différentes sources de données permet d'améliorer la précision des modèles de prédiction agricole [4].

1.5 Système de Prédiction Agricole

Un système de prédiction agricole est un ensemble de processus permettant de transformer des données brutes en estimations de rendement. Il comprend la collecte des données, leur nettoyage, leur prétraitement et leur analyse à l'aide de modèles statistiques ou d'intelligence artificielle [5]. Les résultats sont ensuite présentés sous forme de tableaux ou de visualisations pour aider à la

prise de décision.

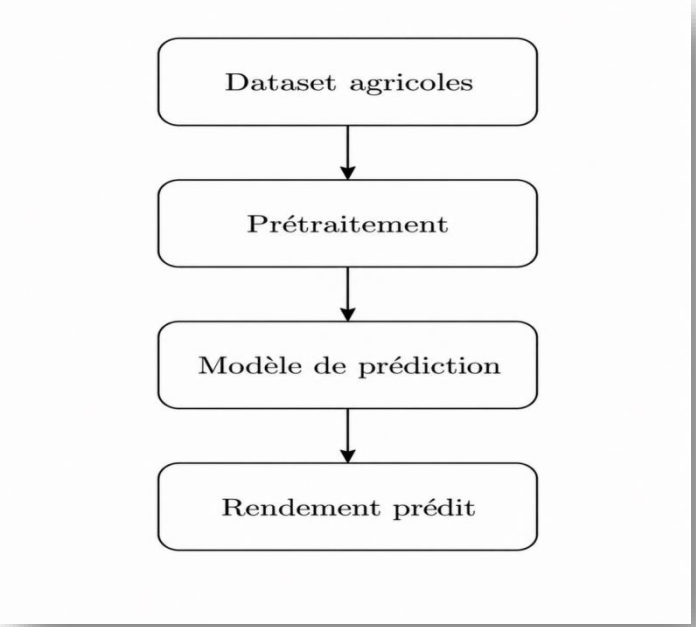


Figure 1: Système de Prédiction Agricole

1.6 Applications des Systèmes de Prédiction Agricole

Les systèmes de prédiction agricole trouvent aujourd'hui de nombreuses applications dans le secteur agricole. Ils sont principalement utilisés pour estimer les rendements des cultures avant la récolte, ce qui permet aux agriculteurs et aux responsables agricoles d'anticiper les niveaux de production et de mieux planifier les activités agricoles.

Ces systèmes contribuent également à la surveillance des cultures en fournissant des informations régulières sur leur état de développement. Grâce à l'analyse des données climatiques et satellitaires, ils permettent d'identifier les zones présentant un risque de stress hydrique ou de faible croissance et d'anticiper les effets des conditions météorologiques sur la production agricole.

Par ailleurs, les prédictions obtenues facilitent une gestion plus efficace des ressources agricoles, notamment l'eau d'irrigation, les engrais et les produits phytosanitaires. Cette optimisation permet de réduire les coûts de production tout en améliorant la productivité des exploitations [1].

Enfin, les systèmes de prédiction peuvent être utilisés pour détecter précocement certaines anomalies ou situations défavorables susceptibles d'affecter les cultures. Les informations fournies aident ainsi les agriculteurs à prendre des mesures correctives au bon moment et à limiter les pertes de rendement [4].

1.7 Défis et Perspectives

Malgré les progrès réalisés dans le domaine de l'agriculture intelligente, plusieurs défis demeurent. La qualité et la disponibilité des données constituent encore une contrainte importante dans de nombreuses régions. La présence de données manquantes, d'erreurs de mesure ou de sources hétérogènes peut affecter la précision des modèles de prédiction [5].

Par ailleurs, les changements climatiques rendent les prévisions agricoles plus complexes. Les variations de température, les sécheresses et les événements météorologiques extrêmes influencent directement la croissance des cultures et augmentent l'incertitude des prédictions [2].

Face à ces défis, les perspectives futures reposent sur l'amélioration des techniques de collecte et d'intégration des données, ainsi que sur le développement de modèles d'intelligence artificielle plus performants. L'utilisation croissante des données satellitaires, des capteurs connectés et des méthodes avancées de Machine Learning et de Deep Learning devrait contribuer à améliorer la précision des systèmes de prédiction agricole et à renforcer leur capacité d'adaptation aux conditions environnementales changeantes [8].

1.8 Conclusion

L'agriculture intelligente constitue aujourd'hui une approche innovante permettant de répondre aux principaux défis auxquels est confronté le secteur agricole. En s'appuyant sur les technologies numériques et l'exploitation des données, elle contribue à améliorer la gestion des exploitations, à optimiser l'utilisation des ressources et à accroître la productivité agricole.

Dans ce contexte, les systèmes de prédiction occupent une place essentielle en fournissant des estimations fiables qui facilitent la prise de décision et la planification des activités agricoles. Grâce à l'intégration de différentes sources de données et à l'utilisation de techniques avancées d'analyse, ces systèmes participent au développement d'une agriculture plus efficace, durable et adaptée aux enjeux futurs.

Le chapitre suivant sera consacré à l'étude des principales méthodes et techniques utilisées pour la prédiction du rendement agricole, avec un intérêt particulier pour les approches basées sur le Machine Learning et le Deep Learning.

Chapitre 2

Intelligence Artificielle et Télédétection pour l'Agriculture Intelligente

2.1 Introduction

Face aux défis croissants liés aux changements climatiques, à la croissance démographique mondiale et aux pressions sur les ressources naturelles, l'agriculture intelligente s'impose aujourd'hui comme une réponse stratégique essentielle. Au cœur de cette transformation, la prédiction du rendement agricole constitue un enjeu majeur, permettant d'anticiper la production, d'optimiser les ressources et de réduire les risques agro-climatiques.

Cependant, cette prédiction reste une tâche complexe en raison de la multiplicité des facteurs influents. Elle nécessite l'intégration de données hétérogènes issues de la télédétection satellitaire, de la météorologie et de l'agronomie, ainsi que le recours à des méthodes d'intelligence artificielle capables de modéliser des relations non linéaires et des dynamiques temporelles complexes.

Dans ce travail, l'étude se concentre exclusivement sur les cultures céréalières stratégiques en Algérie, à savoir le blé dur, le blé tendre, l'orge et l'avoine, qui constituent les principales cultures des Hauts Plateaux et un pilier essentiel de la sécurité alimentaire nationale.

Ce chapitre présente une revue critique et structurée de la littérature scientifique autour de plusieurs axes, notamment l'importance des céréales en Algérie, la télédétection comme source de données, les indices de végétation, ainsi que les techniques de Machine Learning et de Deep Learning. Il se conclut par une synthèse mettant en évidence les principales lacunes scientifiques que cette étude vise à combler et à enrichir.

2.2 Contexte et Enjeux des Céréales en Algérie

2.2.1 Agriculture Intelligente et Prédiction du Rendement

L'agriculture intelligente représente une évolution majeure des pratiques agricoles traditionnelles. Elle repose sur l'intégration de technologies avancées telles que l'Internet des objets (IoT), l'imagerie satellitaire et l'intelligence artificielle, dans le but d'optimiser la production agricole et d'améliorer la durabilité des systèmes céréaliers [7, 9].

Son objectif principal est de collecter, traiter et analyser des données hétérogènes climatiques, édaphiques et agronomiques afin de soutenir la prise de décision [10]. Cette démarche permet d'optimiser les opérations agricoles clés telles que l'irrigation, la fertilisation et la gestion des maladies, tout en réduisant les coûts de production [11]. Toutefois, son adoption reste encore limitée dans plusieurs contextes, en particulier dans les pays en développement, en raison du manque de données fiables, de l'insuffisance des infrastructures numériques et du coût élevé des technologies [12].

La prédiction du rendement des cultures (Crop Yield Prediction – CYP) est définie comme le processus d'estimation de la production potentielle d'une culture donnée à partir de l'analyse de multiples variables : conditions climatiques, propriétés du sol, caractéristiques agronomiques et indices de végétation dérivés de l'imagerie satellitaire [13]. Le rendement est généralement

exprimé en tonnes par hectare.

Selon les projections des Nations Unies, la population mondiale devrait atteindre 9,7 milliards d'habitants d'ici 2050, nécessitant une augmentation de la production alimentaire estimée entre 60 % et 70 % par rapport aux niveaux de 2012 [14]. Pour les céréales en particulier, cette pression est d'autant plus forte que le blé et l'orge constituent des aliments de base pour plus de 2,5 milliards de personnes. Les modèles de prédiction du rendement céréalier sont ainsi devenus des outils stratégiques pour la planification agricole nationale et internationale [13].

2.2.2 Les Céréales en Algérie : Enjeux Stratégiques

Les cultures céréalières — blé dur, blé tendre, orge et avoine — constituent la base de la sécurité alimentaire en Algérie. L'Algérie reste l'un des plus grands importateurs mondiaux de céréales, ce qui rend la prédiction précise des rendements nationaux un enjeu stratégique de premier ordre [8].

Les wilayas des Hauts Plateaux (Sétif, Tiaret, M'Sila, Bordj Bou Arréridj) concentrent l'essentiel de la production céréalière nationale. Cependant, ces zones sont fortement exposées à la variabilité climatique — sécheresses récurrentes, irrégularité des précipitations — qui entraîne des fluctuations importantes des rendements d'une année à l'autre [8]. Dans ce contexte, les modèles de prédiction permettent d'anticiper les risques, d'orienter les politiques d'importation et d'optimiser l'allocation des intrants agricoles.

Les approches de prédiction ont connu une évolution significative. Les modèles statistiques traditionnels (régression linéaire, ARIMA) ont progressivement été remplacés par des méthodes avancées basées sur le Machine Learning (Random Forest, XGBoost, LightGBM) et le Deep Learning (LSTM, CNN, Transformers) [10, 13]. Les études récentes montrent que ces approches modernes offrent une précision prédictive supérieure pour le blé et l'orge, avec des coefficients de détermination pouvant atteindre $R^2 = 0,85–0,93$, contre $R^2 = 0,60–0,75$ pour les modèles classiques [13, 15].

Cependant, leur efficacité reste fortement dépendante de la qualité et de la représentativité des données d'entraînement [13, 15].

2.3 Télédétection : Source de Données Fondamentale pour les Céréales

2.3.1 Principes et Principaux Satellites Utilisés

La télédétection désigne l'ensemble des techniques permettant d'observer et de mesurer les propriétés d'une zone sans contact direct [16]. Elle repose sur des capteurs embarqués sur des satellites, capables d'enregistrer le rayonnement électromagnétique réfléchi ou émis par la surface terrestre [16, 17]. Dans le domaine céréalier, ces données sont largement exploitées pour analyser l'état de santé des cultures, la couverture végétale, l'humidité des sols ainsi que les stress biotiques et abiotiques [18, 19].

Une revue systématique a confirmé que la télédétection constitue aujourd'hui un pilier fondamental de l'estimation des rendements à différentes échelles — locale, régionale et globale — pour le blé, l'orge et les autres céréales [18, 19]. Plusieurs missions satellitaires sont utilisées pour le suivi des céréales :

Sentinel-2 : Développé dans le cadre du programme Copernicus de l'Union européenne, Sentinel-2 offre une résolution spatiale de 10 à 60 mètres avec une revisite de 5 jours. Ses 13 bandes spectrales couvrent le visible, le proche infrarouge (NIR) ainsi que les bandes Red Edge, ce qui le rend particulièrement adapté au calcul des indices de végétation [20]. Il est largement utilisé pour la prédiction du rendement du blé dur, du blé tendre et de l'orge [15, 20].

Landsat : Géré conjointement par la NASA et l'USGS, Landsat fournit des données depuis 1972 avec une résolution spatiale de 30 mètres et une revisite de 16 jours [21]. Sa continuité historique de plus de 50 ans constitue un atout majeur pour les analyses à long terme des tendances de rendement céréalier [21].

MODIS : Les capteurs MODIS fournissent des images avec une résolution spatiale de 250 à 1000 mètres et une revisite quasi quotidienne. Malgré une résolution spatiale plus faible, leur fréquence temporelle élevée permet un suivi dynamique des cycles céréaliers à grande échelle [20]. Une étude comparative récente a montré que Sentinel-2 offre généralement une meilleure précision dans la prédiction du rendement céréalier, tandis que MODIS reste pertinent pour les analyses régionales [20].

2.3.2 Données Radar (Sentinel-1) et Fusion Multi-Sources

Les capteurs radar à synthèse d'ouverture (SAR), tels que Sentinel-1, permettent d'acquérir des données indépendamment des conditions météorologiques et de la couverture nuageuse [22]. Ils opèrent dans le domaine des micro-ondes (bande C) et sont particulièrement sensibles à la structure géométrique des cultures céréalières ainsi qu'à l'humidité des sols [23, 24]. Les polarisations VV et VH fournissent des informations complémentaires sur la structure interne des céréales ; le rapport VH/VV est largement utilisé pour analyser les stades phénologiques du blé et de l'orge, de la levée à la maturité [24, 25].

La fusion multi-sources vise à intégrer ces différentes informations afin d'améliorer la robustesse et la précision des modèles céréaliers [20, 26]. Un modèle XGBoost combinant les données SAR de Sentinel-1, les indices optiques de Sentinel-2 et les données climatiques ERA5-Land a atteint $R^2 = 0{,}81$ pour la prédiction du rendement du blé, trois mois avant la récolte, au Maroc — contexte climatique semi-aride proche de celui des Hauts Plateaux algériens [26].

Dans ce travail, la fusion multi-sources est implémentée au niveau des caractéristiques (feature-level fusion) : les variables issues des différentes sources sont concaténées en un vecteur d'entrée unique fourni aux modèles de prédiction.

2.4 Indices de Végétation pour le Suivi des Céréales

Les indices de végétation sont des combinaisons mathématiques de réflectances spectrales permettant de caractériser l'état de la végétation céréalière. Ils jouent un rôle fondamental dans l'estimation du rendement, car ils fournissent des informations essentielles sur la biomasse, la surface foliaire et l'activité photosynthétique du blé, de l'orge et de l'avoine [18, 27].

NDVI (Normalized Difference Vegetation Index) : L'indice de végétation le plus couramment utilisé en suivi céréalier [20], défini par NDVI = (NIR − R) / (NIR + R). Ses valeurs élevées (0,6–0,9) indiquent une végétation dense et saine, fortement corrélée à la biomasse et au rendement final du blé et de l'orge [15, 18, 27].

EVI (Enhanced Vegetation Index) : Développé pour corriger les limitations du NDVI, notamment la saturation dans les zones à forte biomasse [21]. Il est défini par EVI = G × (NIR − R) / (NIR + C_1R − C_2Blue + L), où G = 2,5, C_1 = 6, C_2 = 7,5 et L = 1. Un modèle Bi-LSTM appliqué au blé d'hiver a identifié l'EVI comme l'une des variables les plus déterminantes pour la prédiction du rendement, notamment en fin de cycle cultural [28].

LAI (Leaf Area Index) : Représente la surface foliaire totale par unité de surface au sol. Directement lié aux processus de photosynthèse et à la production de biomasse des céréales [23], son intégration dans les modèles de prédiction améliore significativement la précision des estimations du rendement du blé et de l'orge [15, 18].

FPAR (Fraction of Absorbed Photosynthetically Active Radiation) : Mesure la fraction de rayonnement photosynthétiquement actif absorbée par la végétation. Il constitue un indicateur direct de la productivité primaire et du potentiel de rendement céréalier [18, 29].

2.5 Modèles de Machine Learning pour la Prédiction du Rendement Céréalier

2.5.1 Vue d'Ensemble

Le Machine Learning s'est imposé comme un outil essentiel dans la prédiction du rendement céréalier. Il permet de transformer des données brutes issues de sources hétérogènes en informations exploitables pour la planification agricole [7, 9]. Cependant, ces modèles présentent certaines limites dans la prise en compte des dépendances temporelles et des relations dynamiques entre variables céréalières [10, 26] .

Dans la suite, nous nous limitons à la présentation des principaux modèles de Machine Learning et de Deep Learning utilisés pour la prédiction du rendement céréalier, qui seront ensuite étudiés et évalués dans le cadre de ce mémoire de master. L'objectif est de mettre en évidence les approches les plus pertinentes afin de les comparer et d'identifier la plus performante pour notre problématique.

2.5.2 Random Forest

Random Forest est l'un des modèles les plus utilisés pour la prédiction du rendement céréalier, en raison de sa robustesse et de sa capacité à traiter des données de grande dimension sans nécessiter

un prétraitement intensif. Cette méthode repose sur la construction d'un ensemble d'arbres de décision dont les prédictions sont agrégées afin d'améliorer la stabilité du modèle [7, 15]. Sa robustesse face aux valeurs aberrantes et aux données manquantes en fait un candidat sérieux pour les zones céréalières algériennes, où les données peuvent être fragmentaires.

2.5.3 Modèles de Boosting : XGBoost et LightGBM

XGBoost est un algorithme de référence pour la prédiction du rendement céréalier [30]. Il repose sur le principe du Gradient Boosting, où les modèles sont construits de manière séquentielle, chaque nouvel apprenant corrigeant les erreurs des précédents. Son efficacité pour le blé est bien documentée : un modèle XGBoost intégrant des données multi-sources (Sentinel-1, Sentinel-2, ERA5-Land) a atteint $R^2 = 0,81$ pour la prédiction du rendement du blé au Maroc, trois mois avant la récolte [26].

LightGBM, développé par Microsoft [31], constitue une alternative efficace pour le traitement de grands ensembles de données céréalières. Il améliore la vitesse d'apprentissage grâce à la technique du Gradient-based One-Side Sampling (GOSS). Son association avec SHAP a démontré sa capacité à modéliser les relations complexes entre variables environnementales et rendement du blé, avec des performances variant entre $R^2 = 0,46$ et 0,66 selon les configurations [32].

2.5.4 Autres Modèles : SVR et CatBoost

Le Support Vector Regressor (SVR) est une extension des machines à vecteurs de support appliquée à la régression. Il cherche à trouver une fonction d'approximation dans un espace de caractéristiques de haute dimension en maximisant la marge de tolérance autour des prédictions. Sa robustesse face aux données de faible volume et aux espaces de grande dimension en fait un candidat pertinent dans les contextes où les données d'entraînement sont limitées, comme c'est le cas pour certaines wilayas céréalières algériennes.

CatBoost, développé par Yandex, est un algorithme de gradient boosting optimisé pour le traitement natif des variables catégorielles, sans nécessiter d'encodage préalable. Il intègre également des mécanismes de régularisation avancés qui limitent le surapprentissage, ce qui le rend adapté aux jeux de données hétérogènes combinant variables climatiques, édaphiques et indices satellitaires.

2.5.5 Limites des Modèles de Machine Learning

Aucun algorithme ne domine de manière universelle dans tous les contextes céréaliers. Les performances varient en fonction de la culture, de la taille des données et des caractéristiques géographiques [11]. Sur le plan des avantages, ces modèles traitent efficacement des données tabulaires multi-sources, sont robustes face aux valeurs aberrantes, offrent un certain niveau d'interprétabilité via SHAP et présentent une bonne efficacité computationnelle [10, 14]. Néanmoins, ils peinent à capturer les dépendances temporelles à long terme essentielles dans les

séries temporelles céréalières [11, 29]. C'est précisément pour dépasser ces limitations que les approches de Deep Learning ont été intégrées.

2.6 Deep Learning pour la Prédiction du Rendement Céréalier

2.6.1 Architecture LSTM

Les réseaux Long Short-Term Memory (LSTM) sont particulièrement adaptés à la modélisation des séries temporelles céréalières. Introduits par Hochreiter et Schmidhuber [33] pour résoudre le problème du gradient vanishing rencontré dans les RNN classiques, les LSTM reposent sur des cellules mémoire et des mécanismes de portes — forget gate, input gate, output gate — qui régulent le flux d'information au fil du temps.

Dans le domaine céréalier, les cycles du blé et de l'orge s'étendent d'octobre à juin sur les Hauts Plateaux. Les LSTM capturent l'effet des conditions initiales de croissance et des stress hydriques ou thermiques tout au long de ce cycle, en intégrant l'évolution temporelle du NDVI et de l'EVI [17, 22]. Un modèle Bi-LSTM appliqué à la prédiction du rendement du blé d'hiver a atteint $R^2 = 0,88$, identifiant l'EVI, la température et les précipitations comme variables déterminantes [28].

2.6.2 Architecture GRU

Les réseaux Gated Recurrent Unit (GRU) constituent une variante simplifiée des LSTM. Ils fusionnent les portes forget et input en une seule porte de mise à jour (update gate) et suppriment la cellule mémoire séparée au profit d'un état caché unifié. Cette simplification réduit le coût computationnel tout en maintenant une capacité comparable à modéliser les dépendances temporelles à moyen terme. Dans le contexte céréalier, les GRU représentent une alternative efficace aux LSTM, particulièrement lorsque les séries temporelles disponibles sont de longueur modérée — ce qui correspond souvent à la réalité des données satellitaires sur les Hauts Plateaux algériens.

2.6.3 Architectures Hybrides et Avancées

Afin de surmonter les limites des approches purement temporelles, des modèles hybrides combinant CNN et LSTM ont été proposés pour la prédiction du rendement céréalier [34]. Dans ces architectures, les CNN extraient les caractéristiques spatiales à partir des données d'imagerie (Sentinel-2), tandis que les LSTM assurent la modélisation des dépendances temporelles. Des architectures CNN-LSTM avec mécanisme d'attention ont permis d'atteindre $R^2 = 0,904$ pour la prédiction du rendement du blé [35].

Le modèle TCN–LSTMs, combinant les Temporal Convolutional Networks et les réseaux LSTM, a démontré de bonnes performances avec $R^2 = 0,81$ pour le blé d'hiver, surpassant CNN-LSTM ($R^2 = 0,78$) et Transformer ($R^2 = 0,69$) [36].

2.7 Comparaison ML vs Deep Learning pour les Céréales

Le choix entre ML et Deep Learning dépend de la nature des données céréalières disponibles. Les modèles ML comme XGBoost demeurent très performants pour l'analyse de données tabulaires structurées [30, 31] et peuvent surpasser le Deep Learning dans certains scénarios spécifiques [37].

Cependant, les architectures profondes présentent un avantage significatif pour traiter des séries temporelles satellitaires couvrant l'ensemble du cycle du blé ou de l'orge [25, 36].Par ailleurs, certains modèles peuvent afficher de très bonnes performances sur des jeux de test classiques mais échouer sur des données temporellement indépendantes ou issues d'un autre contexte géographique [37] — problème particulièrement pertinent pour le contexte algérien, où les données historiques sont limitées.

2.8 Intelligence Artificielle Explicable (XAI) pour les Céréales

Les performances prédictives élevées ne suffisent pas à assurer **l'adoption** des modèles dans le domainecéréalier. Les utilisateurs finaux — agriculteurs, agronomes et décideurs — doivent comprendre les raisons derrière une prédiction afin de lui faire confiance et d'adapter leurs pratiques [32, 38]. C'est dans ce contexte qu'est née l'intelligence artificielle explicable (XAI), qui regroupe des techniques visant à interpréter les décisions des modèles et à identifier les variables les plus influentes sur le rendement céréalier.

Parmi les méthodes XAI, SHAP (Shapley Additive Explanations) est l'approche la plus utilisée. Basée sur la théorie des jeux coopératifs, elle attribue à chaque variable une contribution quantitative à la prédiction finale, permettant une interprétation à la fois globale et locale. Pour la prédiction du rendement du blé, l'association LightGBM + SHAP a permis d'identifier l'importance des indices de végétation en milieu de cycle ainsi que des précipitations tardives comme facteurs clés [32].

Un modèle Bi-LSTM explicable appliqué au blé d'hiver a atteint $R^2 = 0,88$, montrant que l'EVI, la température et les précipitations en fin de cycle jouent un rôle déterminant [28]. Ces résultats sont agronomiquement cohérents avec le comportement du blé et de l'orge dans les zones semi-arides, où le stress hydrique en période de remplissage du grain constitue le principal facteur limitant du rendement.

Une limite importante : une bonne performance sur des ensembles de test standards n'implique pas nécessairement une bonne généralisation temporelle. Certains modèles peuvent produire des explications SHAP cohérentes même en échouant à généraliser sur des données indépendantes [37]. Il est donc essentiel de compléter les analyses XAI par des protocoles de validation rigoureux.
Dans ce travail, SHAP sera appliqué aux modèles les plus performants — meilleur modèle ML et meilleur modèle DL — afin d'identifier les variables climatiques, satellitaires et géospatiales les

plus déterminantes pour le rendement des céréales algériennes, de vérifier la cohérence agronomique des résultats et de fournir des explications interprétables aux utilisateurs finaux via l'application développée.

2.8.1 Tableau Comparatif des Travaux Récents

Le tableau suivant synthétise les principales études recensées dans cet état de l'art, en mettant en évidence les données utilisées, les modèles proposés, les performances obtenues et les limites identifiées.

Tableau 1: Comparaison des principales études sur la prédiction du rendement agricole par IA et Télédétection

Référence	Données	Modèle	R^2	Limite principale
Amazirh et al. [27]	S1 + S2 + ERA5 (blé, Maroc)	XGBoost	0,81	Pas d'explicabilité ; contexte géographique limité
Joshi et al. [20]	NDVI/EVI + Météo (blé d'hiver)	Bi-LSTM + SHAP	0,88	Pas de fusion radar-optique
Song et al. [36]	Sentinel-2 (blé)	CNN-LSTM + Attention	0,904	Complexité élevée ; XAI non intégrée
Wu et al. [37]	Multi-sources (blé d'hiver)	TCN-LSTMs	0,81	Comparaison ML vs DL limitée
Wang et al. [26]	Optique + météo (blé)	LightGBM + SHAP	0,46–0,66	Pas de données radar
Yenkikar et al. [39]	Optique + climat	RF+LSTM+XGBoost+SHAP	0,9827	Pas de données radar ; généralisation limitée
Bouregaa [15]	Données agronomiques (Algérie)	ML classiques	—	Pas de télédétection ; pas de XAI

2.9 Synthèse, Lacunes et Apport du Travail

Principales avancées

L'intégration de données multi-sources (climat, optique, radar) améliore significativement la précision des modèles de prédiction céréalière [7, 9]. Les indices de végétation (NDVI, EVI, LAI, FPAR) sont des indicateurs fiables de la biomasse et de la productivité du blé et de l'orge [18, 28]. Les modèles ML (Random Forest, XGBoost) excellent sur données tabulaires, tandis que les architectures DL (LSTM, CNN-LSTM) capturent efficacement les dépendances temporelles [10, 13]. Les approches XAI (SHAP) sont essentielles pour l'interprétabilité et la confiance [28, 38].

Lacunes identifiées

- Peu d'études comparent rigoureusement XGBoost et LSTM sur un même jeu de données multi-sources céréalières [37].
- La fusion optique-radar-climatique pour les céréales reste partiellement explorée [29].
- L'interprétabilité des modèles de Deep Learning, en particulier des LSTM, demeure limitée dans les applications céréalières réelles [37].
- Le contexte agricole algérien (Hauts Plateaux) est sous-représenté dans la littérature [9],

limitant la transférabilité des modèles.

• Peu de solutions complètes intègrent modélisation, interprétabilité et déploiement opérationnel pour les céréales [38].

Apport du présent travail

Pour répondre à ces lacunes, ce travail propose :

• Une comparaison systématique entre six modèles de Machine Learning (Random Forest, XGBoost, LightGBM, SVR, Gradient Boosting Regressor, CatBoost) et trois architectures de Deep Learning (LSTM, MLP, GRU) appliquées à la prédiction du rendement des céréales algériennes (blé dur, blé tendre, orge, avoine).

• Une stratégie de fusion multi-sources combinant indices optiques (NDVI, EVI, LAI, FPAR), données radar (VV, VH) et variables climatiques.

• L'intégration de SHAP aux modèles les plus performants pour identifier les variables déterminantes, vérifier la cohérence agronomique et fournir des explications aux utilisateurs finaux.

• Le développement d'une application web interactive (Streamlit/Flask) pour la prédiction, la visualisation et l'explication en temps réel du rendement céréalier.

2. 10 Conclusion

Ce chapitre a dressé un état de l'art structuré des approches de prédiction du rendement des cultures céréalières basées sur l'IA et la télédétection. Partant des enjeux stratégiques du blé dur, blé tendre, orge et avoine en Algérie, il a mis en évidence le rôle central des données satellitaires, avant de présenter les principales familles d'algorithmes du Machine Learning classique aux architectures profondes et de souligner l'importance croissante de l'interprétabilité via SHAP.

La revue a révélé plusieurs lacunes persistantes, notamment le manque de comparaisons rigoureuses entre XGBoost et LSTM sur des données céréalières multi-sources, l'exploration limitée de la fusion des données optiques, radar et climatiques pour le blé et l'orge, ainsi que la sous-représentation du contexte agricole algérien [8]. Ces lacunes justifient pleinement l'approche d'intelligence artificielle proposée dans ce mémoire. Le chapitre suivant présentera le jeu de données utilisé, son analyse exploratoire, ainsi que les principales difficultés rencontrées — telles que les valeurs manquantes, l'hétérogénéité et les corrélations — qui motivent le recours aux méthodes avancées de modélisation présentées dans ce travail.

Chapitre 3

Approche hybride proposée

(LSTM+LightGBM+CatBoost)

3.1 Introduction

Ce chapitre est consacré à la présentation de la méthodologie adoptée pour le développement de notre système de prédiction du rendement agricole.

L'objectif est de fournir une vue d'ensemble claire et structurée des différentes étapes suivies pour concevoir une solution basée sur l'intelligence artificielle, capable d'exploiter efficacement les données agricoles, climatiques et satellitaires afin d'estimer le rendement des cultures. Dans un premier temps, nous présenterons l'approche générale proposée ainsi que les différentes sources de données utilisées dans cette étude. Nous détaillerons ensuite les étapes de prétraitement appliquées aux données, notamment le nettoyage, le traitement des valeurs manquantes, la normalisation et l'ingénierie des caractéristiques, suivies d'une analyse exploratoire visant à mieux comprendre la structure des données et les relations existantes entre les différentes variables. Par la suite, nous décrirons les modèles de Machine Learning et de Deep Learning retenus pour la prédiction du rendement agricole, en expliquant leurs principes de fonctionnement ainsi que les raisons de leur sélection.

Enfin, les métriques d'évaluation utilisées pour mesurer et comparer les performances des différents modèles seront présentées afin d'identifier l'approche la plus performante pour la prédiction du rendement agricole.

3.2 Méthodologie de l'approche proposée (hybride LSTM+ LightGBM+CatBoost)

La méthodologie adoptée repose sur une chaîne de traitement structurée comprenant la collecte et le prétraitement des données, suivis par l'analyse exploratoire, l'entraînement de plusieurs modèles de Machine Learning et de Deep Learning, puis la proposition d'un modèle hybride combinant LightGBM, CatBoost et LSTM. Les modèles sont évalués à l'aide de plusieurs métriques de performance et interprétés grâce à la méthode SHAP. Enfin, le meilleur modèle est intégré dans une application web interactive destinée à la prédiction du rendement agricole et à l'aide à la décision.

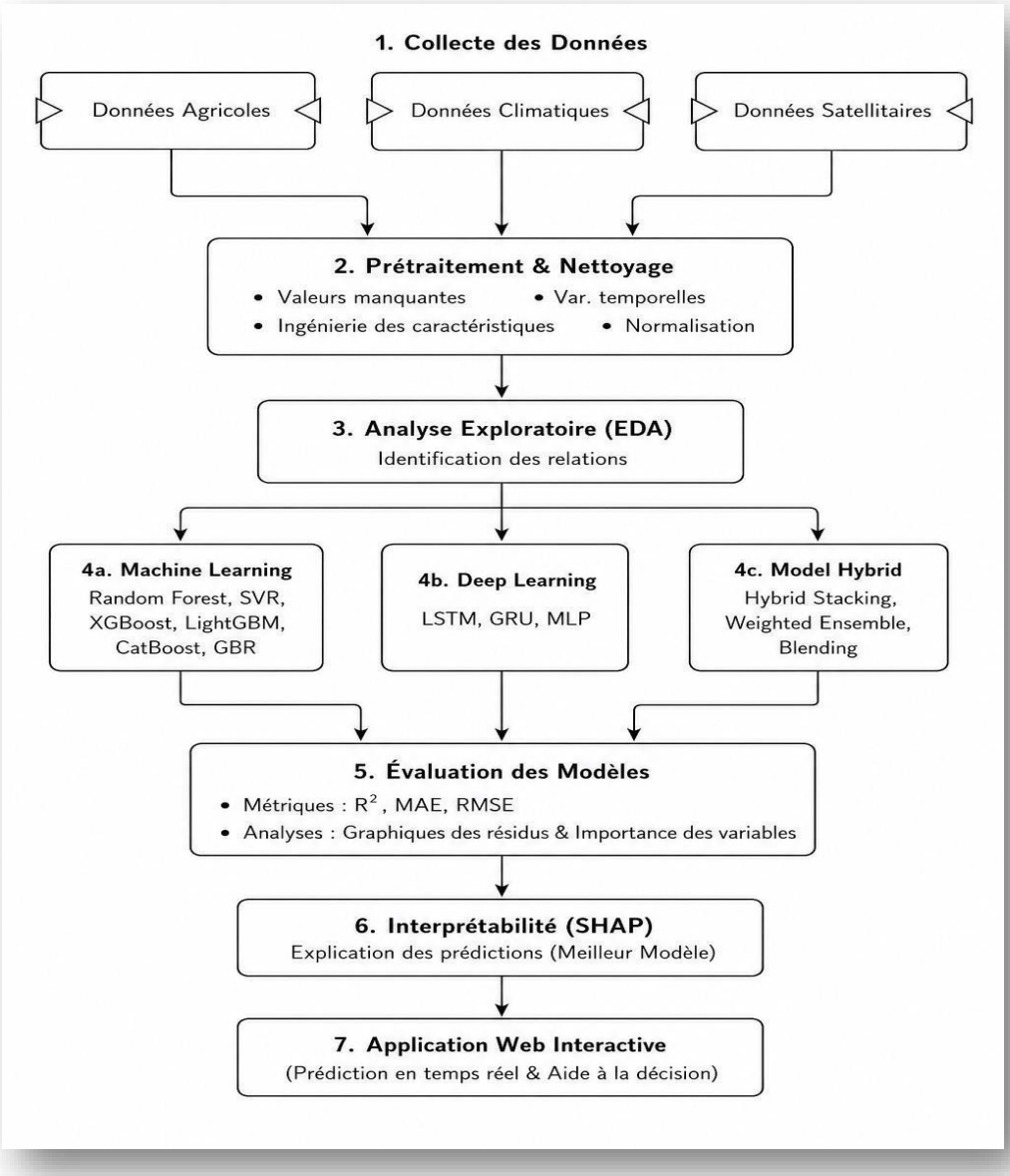


Figure 2: Méthodologie globale de notre système

3.3 Choix du Dataset

Constitue une étape essentielle dans la mise en place d'un système de prédiction du rendement agricole, car la performance des modèles dépend directement de la qualité, de la diversité et de la pertinence des données utilisées. Dans ce travail, nous avons opté pour un dataset multi-sources combinant des informations climatiques, satellitaires, radar et agronomiques afin de représenter de manière réaliste les facteurs influençant la production agricole.

3.3.1 Données climatiques et du sol (Climate & Soil Data)

Les informations climatiques et pédologiques sont issues de la base de données ERA5 -Land [40], qui fournit des données météorologiques à haute résolution spatio-temporelle. Ces données sont particulièrement importantes car elles influencent directement les processus biologiques des cultures ainsi que la disponibilité en eau dans le sol.

Parmi les variables climatiques les plus importantes utilisées dans ce travail, on retrouve :

- la température,
- les précipitations,
- l'humidité du sol,
- l'évaporation.

Ces paramètres jouent un rôle déterminant dans la croissance des plantes et ont un impact direct sur la productivité agricole.

3.3.2 Données radar (Radar Data – Sentinel-1)

Les données radar proviennent du satellite Sentinel-1 [41], qui utilise des capteurs actifs permettant l'observation de la surface terrestre indépendamment des conditions météorologiques et de l'éclairage solaire.

Les coefficients de rétrodiffusion VV et VH constituent les principales variables exploitées. Ils permettent d'analyser la structure du couvert végétal ainsi que le niveau d'humidité du sol. Ces informations sont particulièrement utiles pour suivre l'évolution des cultures dans le temps et détecter les variations de croissance.

3.3.3 Indices de végétation (Vegetation Indices – Sentinel-2)

Les données optiques issues de Sentinel-2 [42] sont utilisées pour le calcul de plusieurs indices de végétation permettant d'évaluer la santé et la densité du couvert végétal.

Les principaux indices utilisés sont :

- NDVI (Normalized Difference Vegetation Index)
- EVI (Enhanced Vegetation Index)
- LAI (Leaf Area Index)
- FPAR (Fraction of Photosynthetically Active Radiation)

Ces indicateurs permettent de quantifier la vigueur des cultures et d'estimer leur potentiel de rendement.

3.3.4 Données de production et variables catégorielles

Les données de rendement agricole ainsi que les variables catégorielles proviennent du Ministère de l'Agriculture et du Développement Rural (MADR) [43], notamment de la Direction des Statistiques Agricoles (Série B, 2016–2019).

Ces données comprennent :

- Le rendement agricole réel (Yield_OfficialYield_Official), qui représente la variable cible du modèle,
- Le type de culture (Crop), incluant les principales céréales cultivées telles que le blé dur, le blé tendre, l'orge et l'avoine,
- La région (Region), couvrant plusieurs wilayas comme Adrar, Biskra, Constantine, Guelma,

Mila et Sétif.

Ces informations permettent d'intégrer le contexte local de production agricole et d'analyser les disparités régionales en termes de rendement.

Tableau 2: Description des attributs du dataset cropyield

N°	Nom de l'attribut	Description de l'attribut	Type
1	ADM1_CODE	Code numérique de la division administrative (Wilaya).	Numérique
2	date	Date d'enregistrement des observations.	Temporel
3	evap_bare_soil	Somme de l'évaporation à la surface du sol nu.	Numérique
4	temperature_2m	Température de l'air à 2 mètres (en Kelvin).	Numérique
5	total_precip	Cumul total des précipitations.	Numérique
6	vol_soil_water_l1	Humidité du sol (couche 1 : 0-7 cm).	Numérique
7	vol_soil_water_l3	Humidité du sol (couche 3 : 28-100 cm).	Numérique
8	year	Année de la campagne de récolte.	Temporel
9	temperature_2c	Variable de température traitée.	Numérique
10	VV	Coefficient radar (polarisation verticale).	Numérique
11	VH	Coefficient radar (polarisation croisée).	Numérique
12	NDVI	Indice de Végétation par Différence Normalisée.	Numérique
13	EVI	Indice de Végétation Amélioré.	Numérique
14	Lai	Indice de Surface Foliaire (Leaf Area Index).	Numérique
15	Fpar	Fraction du rayonnement photosynthétiquement actif.	Numérique
16	Temp_Celsius	Température convertie en degrés Celsius.	Numérique
17	Yield_qx_ha	Valeur brute du rendement.	Numérique
18-23	ADM1_NAME_*	Indicateurs binaires pour les Wilayas (Adrar, Mila, etc).	Binaire
24-27	Crop_*	Indicateurs binaires pour le type de culture.	Binaire
28	Yield_Official	Rendement réel final (Variable Cible).	Numérique

Le Tableau 2 présente l'ensemble des attributs utilisés dans le dataset, en précisant leur nom, leur description et leur type. Ces variables couvrent les données climatiques, satellitaires, radar ainsi que les informations agronomiques et géographiques. Elles incluent également des variables transformées et encodées, notamment les indicateurs binaires des régions et des types de cultures, ainsi que la variable cible représentant le rendement agricole réel.

3.4 Présentation du dataset

Les données sont structurées en un document d'extension (.csv) comportant 90 335 observations et 27 variables. Chaque observation est liée à une combinaison précise de facteurs environnementaux associée à une valeur de rendement agricole exprimée en quintaux par hectare (Yield_Official) Le recours à des données provenant de multiples sources favorise une visualisation plus exhaustive du système agricole et renforce l'habileté des modèles d'apprentissage automatique à saisir les liens complexes entre les variables explicatives et la productivité

	ADM1_CODE	date	evaporation_from_bare_soil_sum	temperature_2m	total_precipitation_sum	volumetric_soil_water_layer_1	volumetric_soil_water_layer_3	year	t
0	341	2016-01-01	-3.250180e-09	287.702880	4.351138e-07	0.000280	-5.158098e-23	2016	
1	341	2016-01-01	-3.250180e-09	287.702880	4.351138e-07	0.000280	-5.158098e-23	2016	
2	341	2016-01-01	-3.250180e-09	287.702880	4.351138e-07	0.000280	-5.158098e-23	2016	
3	341	2016-01-01	-3.250180e-09	287.702880	4.351138e-07	0.000280	-5.158098e-23	2016	
4	341	2016-01-01	-3.250180e-09	287.702880	4.351138e-07	0.000280	-5.158098e-23	2016	
...	...	...	...	...	...	...	...	...	
90331	377	2019-12-30	-2.398459e-04	277.032347	1.919045e-04	0.344529	3.575271e-01	2019	
90332	377	2019-12-30	-2.398459e-04	277.032347	1.919045e-04	0.344529	3.575271e-01	2019	
90333	377	2019-12-30	-2.398459e-04	277.032347	1.919045e-04	0.344529	3.575271e-01	2019	
90334	377	2019-12-30	-2.398459e-04	277.032347	1.919045e-04	0.344529	3.575271e-01	2019	
90335	377	2019-12-30	-2.398459e-04	277.032347	1.919045e-04	0.344529	3.575271e-01	2019	

90336 rows × 27 columns

Figure 3: Capture d'écran de Dataset

3.5 Visualisation de dataset

La visualisation des données est définie comme l'exploration visuelle des données, qui aide à obtenir et connaitre des informations et caractéristiques approfondies et clairs sur le dataset et les variables.

3.5.1 Partitionnement des données

Le nombre d'observations à disposition pour chaque wilaya est présenté dans la figure suivante (Figure 4). Une représentation équilibrée des six wilayas a été notée, chacune offrant approximativement 14 000 observations, ce qui permet une analyse spatiale plus sécurisée de la performance agricole.

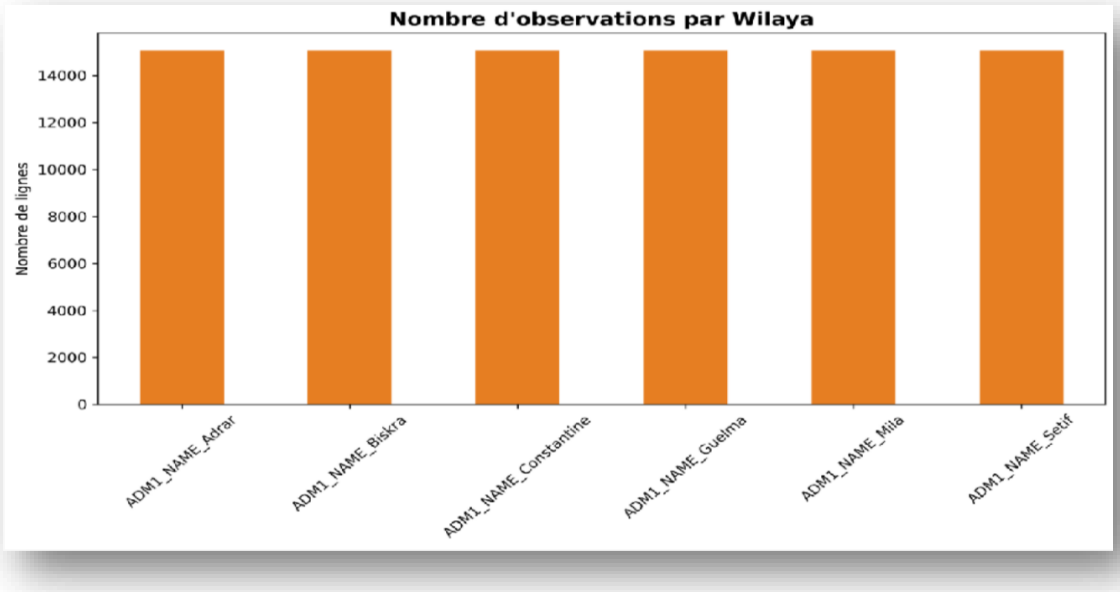


Figure 4: La distribution d'observations par Wilaya

3.5.2 Analyse spatiale et temporelle

La figure 5 présente une analyse spatiale du rendement agricole (Yield_Official) pour les différentes wilayas étudiées à l'aide d'un diagramme en boîte (Boxplot).

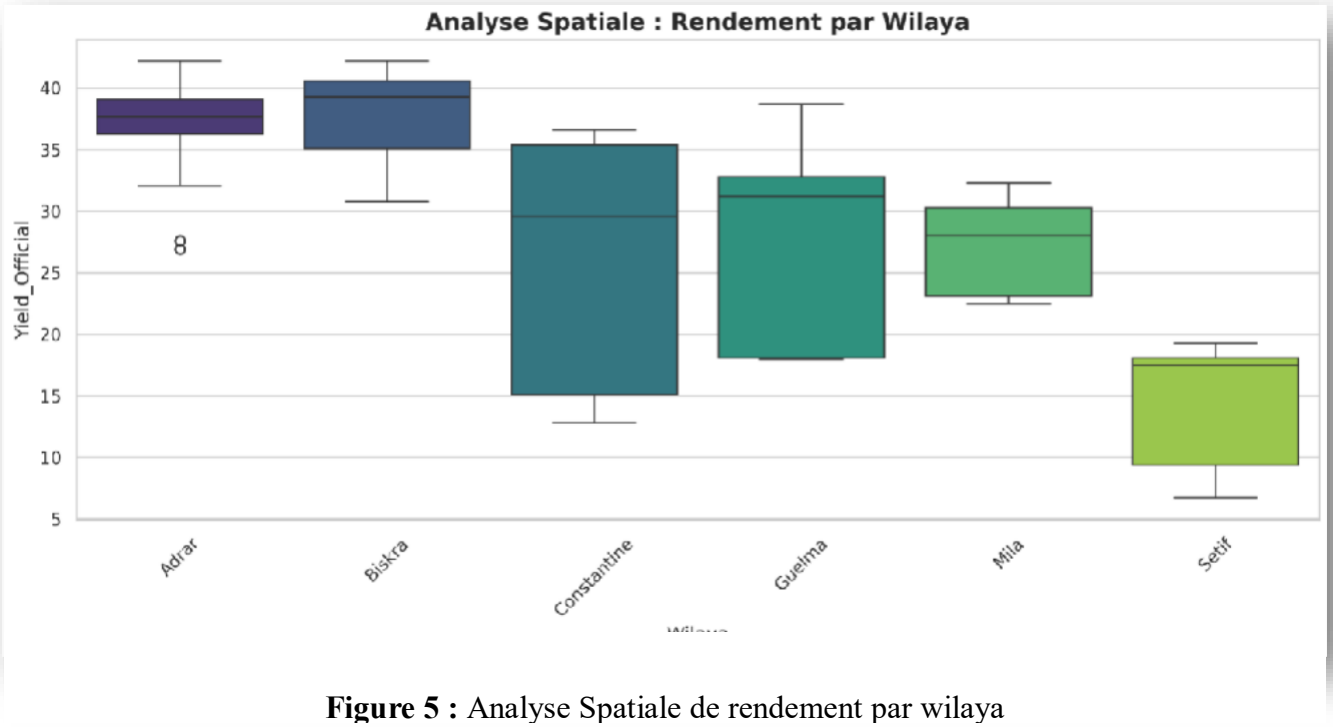


Figure 5 : Analyse Spatiale de rendement par wilaya

Ce type de représentation permet d'évaluer la distribution des rendements, leur dispersion ainsi que les différences entre les régions.

Les résultats montrent que les wilayas de **Biskra** et **Adrar** enregistrent les rendements les plus

élevés, avec des valeurs médianes supérieures à 37 qx/ha. À l'inverse, la wilaya de **Sétif** présente les rendements les plus faibles durant la période étudiée.

Par ailleurs, certaines wilayas telles que **Constantine** et **Guelma** affichent une plus grande variabilité des rendements, ce qui traduit des fluctuations plus importantes d'une année à l'autre. En revanche, **Adrar** et **Biskra** présentent une distribution plus homogène, indiquant une meilleure stabilité de la production agricole.

Ces observations mettent en évidence l'existence de disparités spatiales significatives entre les wilayas étudiées, soulignant l'influence des conditions environnementales et climatiques locales sur le rendement agricole.

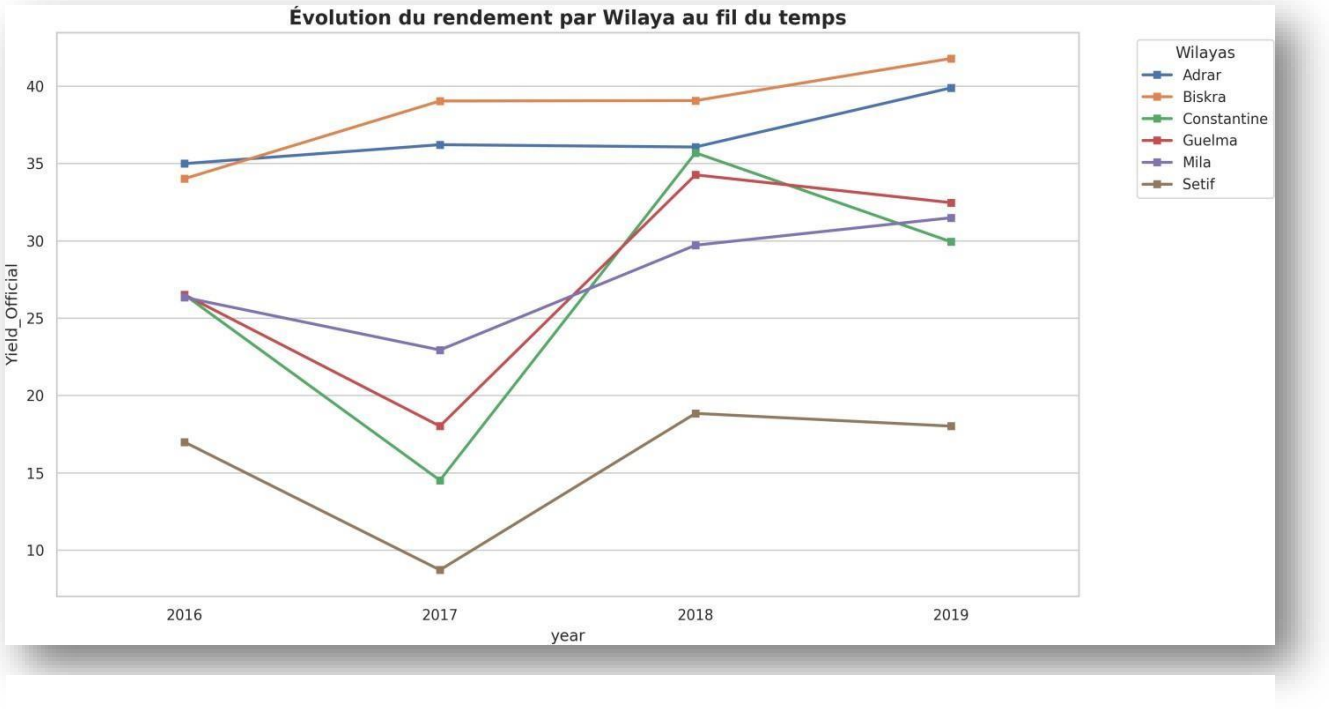


Figure 6 : Évolution du rendement par wilaya

La figure 6 illustre l'évolution du rendement agricole entre 2016 et 2019 dans les différentes wilayas étudiées.

Les résultats montrent une tendance générale à l'amélioration du rendement dans plusieurs wilayas. **Biskra** enregistre les meilleures performances tout au long de la période, avec une augmentation progressive du rendement atteignant plus de 42 qx/ha en 2019. De même, **Adrar** présente une évolution positive et relativement stable au cours des années.

Certaines wilayas, notamment **Constantine** et **Guelma**, connaissent une baisse du rendement en 2017 avant de retrouver des niveaux plus élevés lors des années suivantes. La wilaya de **Mila** affiche quant à elle une progression régulière du rendement durant la période étudiée.

En revanche, **Sétif** demeure la wilaya présentant les rendements les plus faibles malgré une légère

amélioration observée après 2017.

Dans l'ensemble, cette analyse met en évidence l'existence de variations temporelles du rendement agricole, confirmant l'influence des facteurs climatiques, environnementaux et agronomiques sur la production agricole.

3.5.3 Analyse statistique et corrélation

Une analyse exploratoire des données a été réalisée pour identifier les relations entre les variables et la variable cible (Outcome).

Matrice de Corrélation :

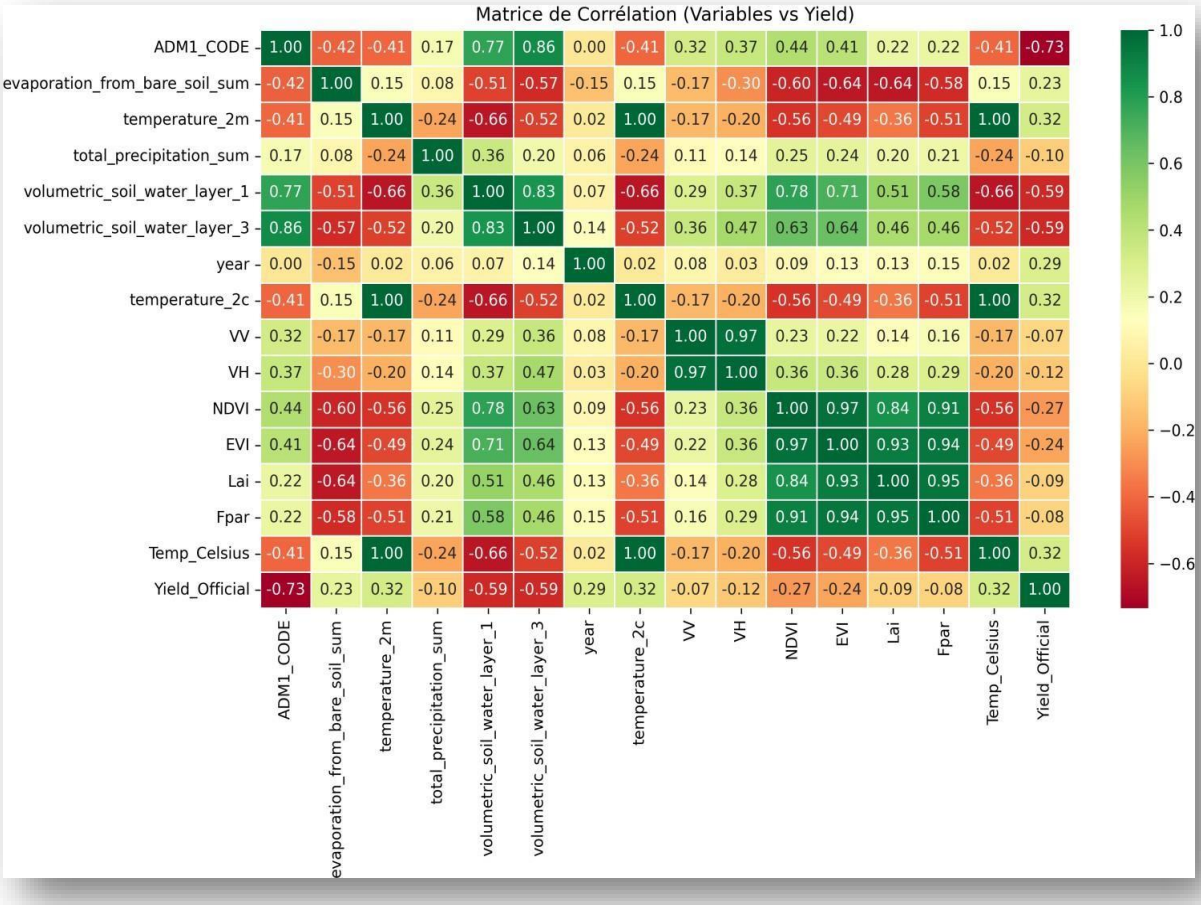


Figure 7 : Matrice de corrélation

3.6 Prétraitement des données

Suite à l'étape d'analyse exploratoire, un processus de prétraitement a été effectué dans le but d'améliorer la qualité des données et de les rendre compatibles avec les modèles d'apprentissage automatique. Cette étape implique diverses manipulations, dont le traitement des valeurs absentes, la modification des variables, la gestion des données liées au temps, la mise à l'échelle ainsi que le fractionnement des données en portions destinées à l'apprentissage et à l'évaluation.

- **Étape 1 : Imputation des valeurs manquantes (Moyenne et Mode)**

Les valeurs manquantes peuvent affecter négativement les performances des modèles

prédictifs. Afin de préserver l'ensemble des observations disponibles, un traitement différencié a été appliqué : les valeurs manquantes des variables numériques ont été remplacées par la moyenne de leur colonne respective (Mean Imputation), tandis que pour les variables catégorielles, la modalité la plus fréquente (Mode Imputation) a été adoptée pour combler les manques.

Python :

```
# Imputation des valeurs manquantes (Moyenne et Mode)
num_cols = df.select_dtypes(include=['number']).columns
df[num_cols] = df[num_cols].fillna(df[num_cols].mean())

cat_cols = df.select_dtypes(include=['object']).columns
for col in cat_cols:
    df[col] = df[col].fillna(df[col].mode()[0])
```

- **Étape 2 : Encodage des types booléens et ingénierie des dates**

Certaines variables du jeu de données étant représentées sous forme logique (True/False), elles ont été converties en valeurs numériques binaires (0 et 1). Parallèlement, la variable de date brute a été décomposée pour capturer la saisonnalité agricole. Plusieurs caractéristiques temporelles en ont été extraites, notamment l'année, le mois et le jour de l'année (de 1 à 365), avant que la colonne de date originale ne soit définitivement supprimée.

Python :

```
# Encodage des types Booléens et Ingénierie des dates
bool_cols = df.select_dtypes(include=['bool']).columns
df[bool_cols] = df[bool_cols].astype(int)

if 'date' in df.columns:
    df['date'] = pd.to_datetime(df['date'], errors='coerce')
    df['year'] = df['date'].dt.year
    df['month'] = df['date'].dt.month
    df['day_of_year'] = df['date'].dt.dayofyear
    df.drop(columns=['date'], inplace=True)
```

- **Étape 3 : Nettoyage de la cible et encodage catégoriel**

La variable cible utilisée dans cette étude correspond au rendement agricole réel (Yield_Official). Les observations présentant des valeurs manquantes ou nulles ont été supprimées pour éviter d'introduire des incohérences. Ensuite, les variables catégorielles restantes ont été transformées en vecteurs numériques via la méthode One-Hot Encoding (pd.get_dummies) avec l'option drop_first=True pour éliminer la redondance et éviter le piège de la multicolonnalité.

Python :

```
# Nettoyage de la cible et Encodage Catégoriel (One-Hot)
df = df[df['Yield_Official'].notna() & (df['Yield_Official'] > 0)]

# Encodage disjonctif complet des variables qualitatives
df = pd.get_dummies(df, columns=df.select_dtypes(include=['object']).columns, drop_first=True)
```

- **Étape 4 : Séparation des caractéristiques et de la cible**

Après la phase de nettoyage et de codage, le jeu de données a été scindé en deux entités mathématiques distinctes : la matrice des descripteurs (X) regroupant les variables explicatives climatiques, satellitaires et agricoles, et le vecteur cible (y) correspondant au rendement (Yield_Official). Les colonnes de contrôle comme l'identifiant administratif (ADM1_CODE) et l'année ont été retirées pour éviter toute fuite d'information (Data Leakage).

Python :

```
# Séparation des caractéristiques (X) et de la cible (y)
years_vector = df['year'] if 'year' in df.columns else df.filter(like='year').iloc[:,0]
y = df['Yield_Official']
X = df.drop(columns=['Yield_Official', 'ADM1_CODE', 'year'], errors='ignore')
```

- **Étape 5 :Répartition temporelle des données**

Contrairement à une division aléatoire classique, une séparation temporelle a été adoptée afin de mieux reproduire un scénario réel de prédiction agricole.

Les données des années **2016 à 2018** ont été utilisées pour l'entraînement, tandis que les données de **2019** ont été réservées pour les tests et l'évaluation des modèles.

Cette stratégie permet de vérifier la capacité des modèles à généraliser sur une campagne agricole future jamais observée durant l'apprentissage

Python :

```
# Répartition temporelle stricte (Train: 2016-2018 / Test: 2019)
X_train = X[years_vector < 2019]
y_train = y[years_vector < 2019]

X_test = X[years_vector == 2019]
y_test = y[years_vector == 2019]
```

- **Étape 6 : Standardisation des données (Centrage-Réduction)**

Les variables numériques présentent des échelles différentes. Afin d'assurer une convergence plus rapide des modèles et d'éviter qu'une variable domine les autres, une standardisation a été appliquée à l'aide de la méthode **StandardScaler**.

Cette transformation centre les données autour de zéro et leur attribue une variance égale à un.

Python :

```
# Standardisation des données (Centrage-Réduction)
scaler = StandardScaler()
X_train_ready = pd.DataFrame(scaler.fit_transform(X_train), columns=X_train.columns)
X_test_ready = pd.DataFrame(scaler.transform(X_test), columns=X_test.columns)
```

3.7 Ingénierie des fonctionnalités

L'ingénierie des variables constitue une phase cruciale dans cette initiative, dont l'objectif est de convertir les données non traitées en variables informatives susceptibles d'optimiser l'efficacité des modèles prédictifs. Dans le cadre de l'agriculture intelligente, cette phase est particulièrement cruciale compte tenu de la variété et de la complexité des données employées (climatiques, satellitaires et agricoles).

Plutôt que d'utiliser directement les données brutes, diverses transformations ont été mises en œuvre dans le but de tirer des informations pertinentes et utilisables par les modèles.

3.7.1 Extraction des indices de végétation et des données radar

Plutôt que d'employer directement les images satellites brutes, des indices ont été élaborés pour illustrer quantitativement l'état de la végétation.

Dans ces indices :

NDVI (Indice de Végétation par Différence Normalisée) et EVI (Indice de Végétation Amé- lioré) : Ces indices proviennent des bandes spectrales, rouge et infrarouge proche, du satellite Sentinel-2. ont pour fonction d'évaluer l'état de santé, la densité et la vitalité de la végétation **VV et VH (informations radar)** : Provenant du satellite Sentinel-1, ces variables symbolisent les coefficients de rétrodiffusion radar. Elles offrent des renseignements sur la composition du couvert végétal et l'humidité du sol.

Cette transformation représente une étape avancée de l'ingénierie des caractéristiques, puisqu'elle offre la possibilité de transmuter des données complexes (telles que les images et les signaux radar) en variables numériques pertinentes pour la prédiction.

3.7.2 Encodage des variables catégorielles

Les données agricoles contiennent des variables qualitatives telles que :

- La région ou wilaya (ex : Mila)
- Le type de culture (ex : Blé dur)

Ces variables ne peuvent pas être exploitées directement par les modèles de Machine Learning. Afin de les rendre utilisables, une transformation a été appliquée à l'aide de techniques d'encodage, notamment :

- One-Hot Encoding **:** transformation des catégories en variables binaires (0/1) Cette approche permet au modèle de traiter chaque catégorie comme une entité distincte et d'ef- fectuer des comparaisons numériques entre différentes cultures et régions.

3.7.3 Encodage des variables temporelles

Les données comportent un aspect temporel (dates d'observation). Pour tirer parti de ces informa- tions, certaines variables temporelles ont été isolées, y compris :

- Année (Year)

- Mois (Month)
- Jour (Day)

3.7.4 Intégration des données provenant de plusieurs sources (Fusion de don- nées)

Une des contributions significatives de ce projet est l'intégration de données issues de diverses sources :

- Données météorologiques (ERA5-Land)
- Données issues des satellites (Sentinel-1 et Sentinel-2)
- Informations agricoles (MADR)

Ces données ont été consolidées en une structure unique et cohérente, où chaque ligne illustre un assemblage précis de conditions météorologiques, d'indices de végétation et de traits agricoles pour une région spécifique.

Cette méthode facilite l'établissement d'un lien direct entre des variables comme :

- L'humidité du sol et les précipitations
- L'indice de végétation (NDVI)
- La productivité agricole

L'intégration des données représente une phase avancée du L'ingénierie des variables, car elle offre la possibilité de tirer parti des interactions entre diverses sources d'information, augmentant ainsi l'ap- titude des modèles à saisir la complexité du système agricole.

3.8 Sélection des variables

La sélection des variables constitue une étape importante dans le développement d'un modèle de prédiction du rendement agricole. Elle permet de réduire la complexité des modèles, de limiter le risque de surapprentissage (Overfitting) et d'améliorer les performances ainsi que l'interprétabilité des résultats.

Après l'analyse exploratoire des données, l'étude de la matrice de corrélation et les différentes étapes d'ingénierie des caractéristiques, les variables retenues regroupent des informations climatiques, satellitaires et agricoles présentant une influence potentielle sur le rendement agricole. Les caractéristiques finales utilisées pour l'entraînement des modèles sont :

evaporation_from_bare_soil_sum,temperature_2m,total_precipitation_sum, volumetric_soil_water_layer_1, volumetric_soil_water_layer_3, temperature_2c, VV, VH, NDVI, EVI, LAI, FPAR, Temp_Celsius, month, day_of_year, ainsi que les variables catégorielles issues de l'encodage des cultures et des wilayas.

Cette sélection permet de conserver les informations les plus pertinentes pour la prédiction du rendement agricole tout en réduisant les redondances entre les variables explicatives.

3.9 Paramétrage des Algorithmes

Le paramétrage des algorithmes consiste à configurer les hyperparamètres des modèles d'apprentissage automatique afin d'obtenir les meilleures performances possibles. Dans ce travail, les paramètres des différents modèles ont été définis manuellement sur la base de la littérature scientifique et de plusieurs essais expérimentaux réalisés sur le jeu de données agricole. Cette étape permet d'assurer un bon compromis entre la précision des modèles et leur capacité de généralisation.

Le réglage des hyperparamètres joue un rôle essentiel dans la performance finale des modèles. Plusieurs stratégies ont été mises en œuvre afin d'adapter chaque algorithme à la nature des données utilisées.

3.9.1 Hyperparamètres spécifiques pour chaque modèle

Les principaux paramètres utilisés pour les modèles d'apprentissage automatique et de Deep Learning sont les suivants :

- **Nombre d'arbres (n_estimators)** : contrôle la complexité des modèles basés sur les arbres tels que Random Forest, LightGBM et CatBoost.
- **Taux d'apprentissage (learning_rate)** : permet de réguler la vitesse d'apprentissage dans les modèles de boosting.
- **Random state / random seed** : assure la reproductibilité des résultats pour tous les modèles.
- **Nombre d'unités LSTM (LSTM units)** : définit la dimension des représentations extraites par le réseau LSTM.
- **Nombre d'unités GRU** : contrôle la capacité du modèle GRU à apprendre les dépendances temporelles.
- **Nombre de neurones dans MLP** : détermine la complexité du réseau de neurones feedforward.
- **Batch size et nombre d'époques** : influencent la stabilité et la durée de l'entraînement des modèles de Deep Learning.
- **Fonctions d'activation (ReLU, tanh)** : utilisées pour introduire la non-linéarité dans les réseaux neuronaux.

3.9.2 Validation et réglage des modèles

Dans cette étude, les modèles ont été évalués à l'aide d'un ensemble de validation (validation split) lors de l'entraînement des réseaux de neurones. Cette approche permet de surveiller la performance du modèle sur des données non vues et de limiter le surapprentissage.

Pour les modèles de Deep Learning, un mécanisme d'**Early Stopping** a été utilisé afin d'arrêter automatiquement l'entraînement lorsque la performance sur les données de validation ne s'améliore plus, ce qui permet d'éviter le surapprentissage et d'améliorer la généralisation du

modèle

3.9.3 Gestion de l'apprentissage et optimisation des modèles

L'optimisation des modèles a été assurée par :

- L'utilisation de l'optimiseur **Adam** pour les réseaux de neurones (LSTM, GRU et MLP), permettant une convergence rapide et stable.
- L'utilisation de paramètres fixes pour les modèles de Machine Learning (LightGBM et CatBoost) afin de garantir un comportement stable et reproductible.
- La combinaison de plusieurs modèles dans l'approche hybride afin d'améliorer la robustesse des prédictions.
- L'utilisation de la méthode **Ensemble Averaging** pour fusionner les sorties des modèles LightGBM et CatBoost dans le modèle hybride.
- L'interprétation des résultats à l'aide de la méthode **SHAP**, permettant d'analyser l'importance des variables dans la prédiction finale du rendement agricole.

3.9.4 Métriques d'évaluation

Pour évaluer la performance des modèles de régression, nous avons utilisé les métriques suivantes :

- MSE (Mean Squared Error)

L'erreur quadratique moyenne correspond à la moyenne des carrés des écarts entre les valeurs prédites et les valeurs réellement observées. Elle pénalise davantage les grandes erreurs, ce qui la rend particulièrement sensible aux valeurs aberrantes (outliers) [44].

$$MSE = (1/n) \times \Sigma (y_i - \hat{y}_i)^2$$

- RMSE (Root Mean Square Error)

La racine de l'erreur quadratique moyenne correspond à la racine carrée de la MSE. Exprimée dans la même unité que la variable cible (rendement céréalier), elle facilite l'interprétation directe des résultats [45].

$$RMSE=\sqrt{[(1/n)\times\Sigma(y_i-\hat{y}_i)^2]}$$

- MAE (Mean Absolute Error)

L'erreur absolue moyenne représente la moyenne des écarts absolus entre les valeurs prédites et les valeurs réelles. Contrairement au RMSE, elle traite toutes les erreurs de manière proportionnelle, ce qui la rend plus robuste aux valeurs aberrantes [46].

$$MAE = (1/n) \times \Sigma |y_i - \hat{y}_i|$$

- R^2 (Coefficient de détermination)

Le coefficient de détermination mesure la proportion de la variance de la variable cible expliquée par le modèle. Sa valeur est généralement comprise entre 0 et 1 : plus elle est proche de 1, plus le modèle est performant [47].

$$R^2 = 1 - [\,(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2\,]$$

- MAPE (Mean Absolute Percentage Error)

L'erreur absolue moyenne en pourcentage exprime l'erreur de prédiction en proportion des valeurs réelles. Elle est facilement interprétable, mais devient instable lorsque les valeurs réelles sont proches de zéro [48].

$$MAPE = (100/n) \times \Sigma|y_i - \hat{y}_i| / |y_i|$$

- MedAE (Median Absolute Error)

L'erreur absolue médiane correspond à la médiane des erreurs absolues entre les valeurs réelles et les prédictions. Cette métrique est particulièrement robuste aux valeurs aberrantes et constitue une alternative fiable à la MAE [49].

$$MedAE = mé\,(\,|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|\,)$$

- EVS (Explained Variance Score)

Le score de variance expliquée mesure la proportion de la variance des données réelles expliquée par le modèle. Une valeur proche de 1 indique un bon ajustement [49].

$$EVS = 1 - (y - \hat{y}) / (y)$$

- Adj. R^2 (Coefficient de détermination ajusté)

Le R^2 ajusté est une version corrigée du R^2 qui pénalise l'ajout de variables non pertinentes au modèle. Il est indispensable pour évaluer fidèlement la performance réelle d'un modèle multivarié [46].

$$Adj.R2 = 1 - [\,(1 - R^2)(n - 1) / (n - p - 1)\,]$$

Où :

- **n** : désigne le nombre total d'observations (échantillons).
- **p** : désigne le code/nombre de variables explicatives (*features)*.
- $\mathbf{y_i}$: est la valeur réelle observée.
- $\hat{\mathbf{y}}_\mathbf{i}$: est la valeur prédite par le modèle.
- $\bar{\mathbf{y}}$: représente la moyenne des valeurs réelles.

3.10 Description des modèles Machine learning

Dans ce projet, divers modèles d'apprentissage supervisé ont été employés pour estimer la productivité agricole à partir de données provenant de multiples sources, comprenant les informations climatiques, satellitaires et agricoles. Ces modèles ont subi un entraînement suite à une étape de prétraitement qui a inclus le nettoyage des données, le traitement des valeurs absentes et l'extraction des variables temporelles.

3.10.1 Modèles d'apprentissage supervisé

Random Forest

L'algorithme Random Forest est un algorithme d'apprentissage supervisé basé sur un ensemble d'arbres de décision. Ce modèle a été choisi pour sa robustesse face au bruit, sa capacité à modéliser des relations non linéaires et son efficacité sur les données tabulaires.

Le modèle est configuré avec les paramètres suivants : **n_estimators = 100**, **random_state = 42** et **n_jobs = -1**. Il permet également d'évaluer l'importance des variables explicatives afin d'identifier les facteurs les plus influents dans la prédiction du rendement agricole.

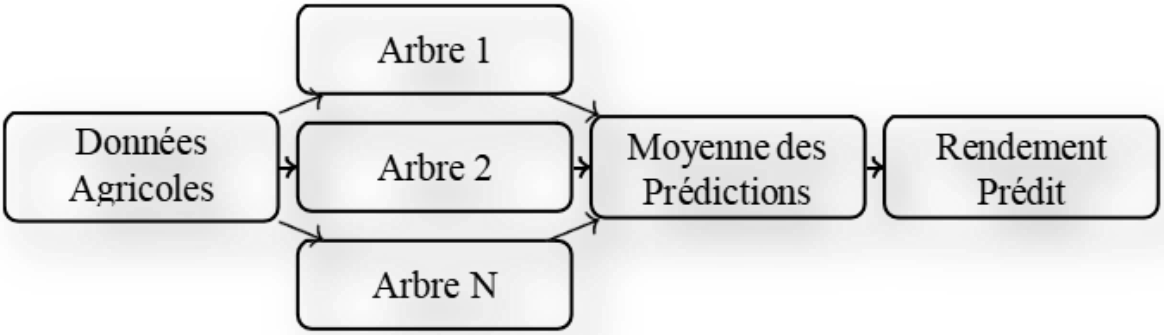


Figure 8: Architecture du Modèle RF

SVR (Support Vector Regression)

SVR est un algorithme d'apprentissage supervisé basé sur les machines à vecteurs de support. Il permet de modéliser des relations non linéaires complexes grâce à l'utilisation de fonctions noyaux (Kernel Functions), en particulier le noyau RBF utilisé dans ce travail.

Le modèle SVR est configuré avec un noyau **RBF (Radial Basis Function)** ainsi que des hyperparamètres ajustés manuellement, notamment **C = 10** pour contrôler le compromis entre précision et régularisation, et **epsilon = 0.1** pour définir la marge de tolérance dans la fonction de perte. Avant l'entraînement, les données sont standardisées afin d'améliorer la stabilité du modèle et garantir de meilleures performances.



Figure 9: Architecture du modèle SVR

3.10.2 Algorithmes de Gradient Boosting

Gradient Boosting Regressor

Gradient Boosting Regressor est un algorithme d'apprentissage supervisé basé sur la technique

du Gradient Boosting. Il construit séquentiellement plusieurs arbres de décision, chaque nouvel arbre étant entraîné pour corriger les erreurs de l'arbre précédent. Cette approche permet d'améliorer progressivement la précision des prédictions et de capturer des relations complexes entre les variables.

Dans ce projet il a été retenu pour sa capacité à fournir des performances élevées sur les données tabulaires et à modéliser efficacement les relations non linéaires.

Le modèle est configuré avec les paramètres suivants : **n_estimators = 200** et **random_state = 42**. Après l'entraînement, les performances sont évaluées à l'aide de différentes métriques ainsi que par une analyse des résidus. L'importance des variables explicatives est également calculée afin d'identifier les facteurs les plus influents dans la prédiction du rendement agricole.

XGBoost

XGBoost est un algorithme d'apprentissage supervisé basé sur la technique du Gradient Boosting. Il construit successivement plusieurs arbres de décision afin de corriger les erreurs des arbres précédents et d'améliorer progressivement la précision des prédictions.

Le modèle est configuré avec les paramètres suivants : **n_estimators = 300**, **learning_rate = 0.05**, **subsample = 1**, **colsample_bytree = 1**, **random_state = 42** et **n_jobs = -1**. Après l'entraînement, les performances du modèle sont évaluées à l'aide de différentes métriques de régression et d'une analyse des résidus. De plus, l'importance des variables explicatives est calculée afin d'identifier les facteurs les plus influents dans la prédiction du rendement agricole.

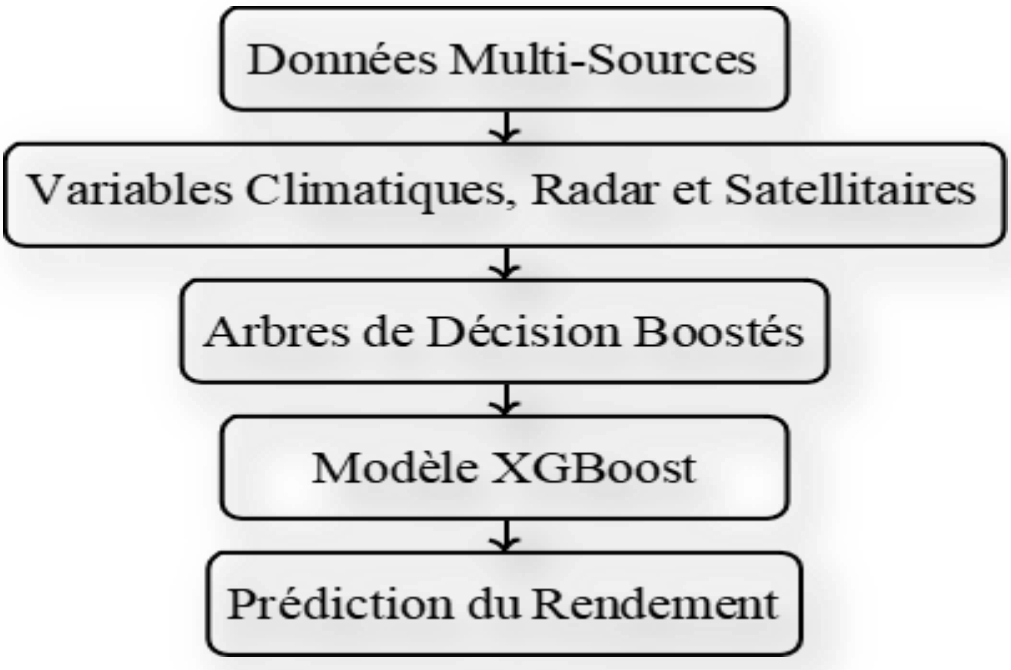


Figure 10: Architecture du Modèle XGBoost

LightGBM

LightGBM est un algorithme d'apprentissage supervisé basé sur la technique du Gradient Boosting. Il est conçu pour offrir une grande rapidité d'exécution et une utilisation efficace de la mémoire grâce à sa stratégie de croissance basée sur les feuilles (Leaf-wise Growth).

Le modèle est configuré avec les paramètres suivants : **n_estimators = 300**, **random_state = 42** et **n_jobs = -1**. Il permet également d'évaluer l'importance des variables explicatives afin d'identifier les facteurs les plus influents dans la prédiction du rendement agricole.

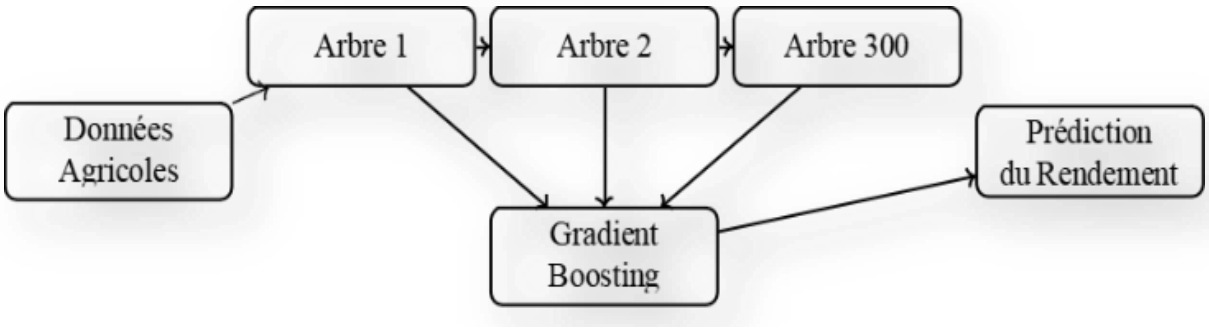


Figure 11 : Architecture du Modèle LightGBM

CatBoost

CatBoost est un algorithme d'apprentissage supervisé basé sur la technique du Gradient Boosting il est conçu pour améliorer les performances prédictives tout en réduisant le risque de surapprentissage grâce à différents mécanismes d'optimisation intégrés.

Le modèle est configuré avec les paramètres suivants : **iterations = 300** et **verbose = 0**. Après l'entraînement, les performances du modèle sont évaluées à l'aide de différentes métriques de régression ainsi que par une analyse des résidus. L'importance des variables explicatives est également étudiée afin d'identifier les facteurs les plus influents dans la prédiction du rendement agricole.

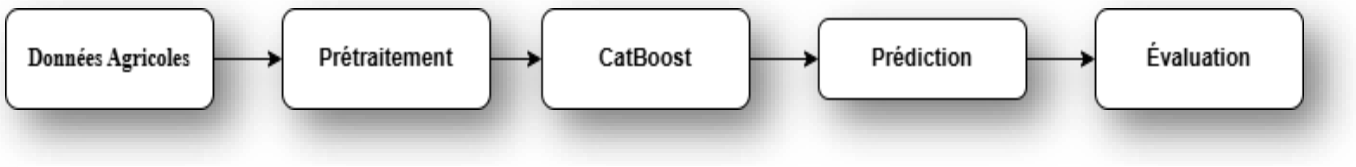


Figure 12 : Architecture du Modèle CatBoot

3.11 Description des modèles deep learning

LSTM (Long Short-Term Memory)

Le modèle LSTM (Long Short-Term Memory) est un réseau de neurones récurrent conçu pour modéliser les relations complexes et non linéaires présentes dans les données. Grâce à sa structure interne à mémoire, il est capable d'extraire des représentations pertinentes à partir des variables d'entrée et d'améliorer la qualité des prédictions.

Les données sont d'abord standardisées puis restructurées sous forme tridimensionnelle afin d'être compatibles avec l'architecture LSTM. Le modèle est composé d'une couche **LSTM de 64 unités**, suivie d'une couche **Dense de 32 neurones** utilisant la fonction d'activation **ReLU**, puis d'une couche de sortie **Dense à un neurone** destinée à la prédiction du rendement agricole.

Le modèle est optimisé avec l'algorithme **Adam** et entraîné avec les paramètres suivants : **epochs = 100**, **batch_size = 64**, **validation_split = 0.2**, **patience = 10** (Early Stopping) et **random_state = 42**. Les performances du modèle sont ensuite évaluées à l'aide de différentes métriques de régression, d'une analyse des résidus ainsi que de la courbe d'apprentissage basée sur l'évolution de la fonction de perte (loss Curve).

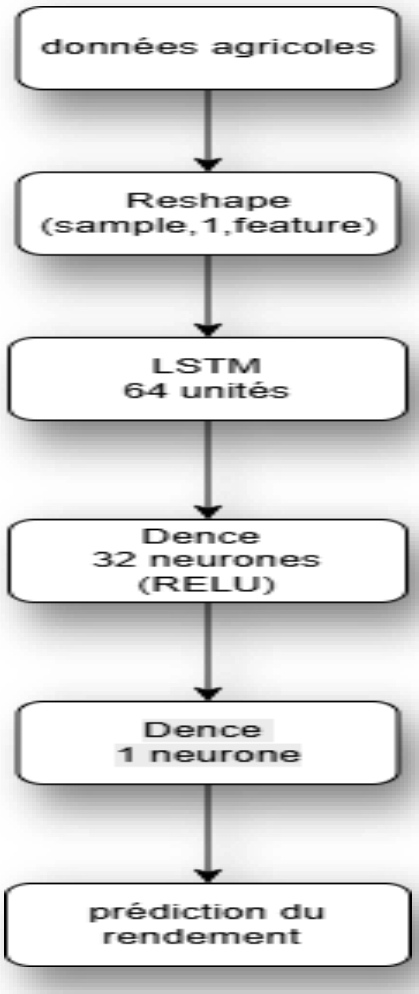


Figure 13 : Architecture du Modèle LSTM

GRU (Gated Recurrent Unit)

Le modèle **GRU (Gated Recurrent Unit)** est une architecture de réseau de neurones récurrente conçue pour le traitement des données séquentielles et temporelles. Il constitue une version simplifiée du modèle LSTM, permettant de capturer les dépendances temporelles tout en réduisant le nombre de paramètres à apprendre.

Avant l'entraînement, les données ont été standardisées à l'aide de **StandardScaler**, puis converties sous forme tridimensionnelle afin d'être compatibles avec l'architecture récurrente du modèle. Le réseau développé est composé d'une couche **GRU de 64 neurones** utilisant la fonction d'activation **tanh**, suivie d'une couche dense de **32 neurones** avec activation **ReLU**, puis d'une couche de sortie contenant un seul neurone pour la prédiction du rendement.

Le modèle a été entraîné avec l'optimiseur **Adam** et la fonction de perte **Mean Squared Error (MSE)** pendant **100 époques**, avec une taille de lot (**batch size**) de **256**. Ce modèle a été retenu pour sa capacité à modéliser efficacement les relations temporelles présentes dans les données agricoles tout en nécessitant un temps d'apprentissage inférieur à celui des réseaux LSTM.

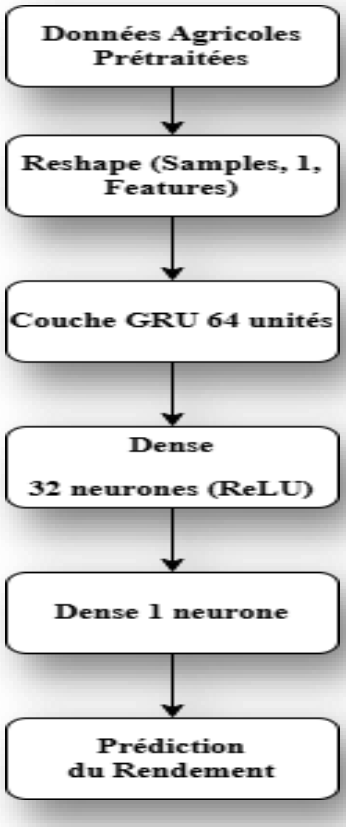


Figure 14 : Architecture du Modèle GRU

MLP (Multi-Layer Perceptron)

Le modèle **MLP (Multi-Layer Perceptron)** est un réseau de neurones artificiels de type feed-forward largement utilisé pour les tâches de régression. Grâce à ses couches cachées et à ses fonctions d'activation non linéaires, le MLP est capable d'apprendre des relations complexes entre les caractéristiques d'entrée et le rendement des cultures.

Le modèle développé est composé de **trois couches cachées** contenant respectivement **128, 64 et 32 neurones**, avec la fonction d'activation **ReLU**. L'apprentissage est réalisé à l'aide de l'optimiseur **Adam** et de la fonction de perte **Mean Squared Error (MSE)**. Les données ont été préalablement standardisées à l'aide de **StandardScaler** afin d'améliorer la stabilité de l'apprentissage. Le modèle a été entraîné pendant **100 époques (epochs)** avec une taille de lot (**batch size**) de **64** et une validation interne de **20 %** des données d'entraînement. Les performances ont ensuite été évaluées sur les données de test afin de comparer son efficacité avec les autres modèles étudiés.

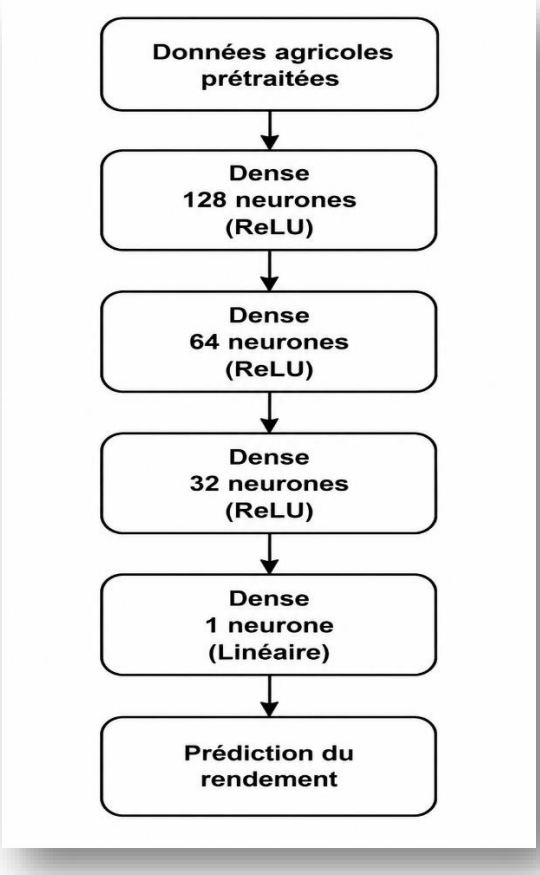


Figure 15 : Architecture du Modèle MLP

3.12 Modéle proposé

Modèle Hybrid (LSTM + LightGBM + CatBoost)

Le modèle hybride proposé combine les avantages du **Deep Learning** et du **Machine Learning** afin d'améliorer la précision de la prédiction du rendement agricole. Cette approche repose sur l'utilisation d'un réseau **LSTM** pour extraire des représentations avancées des données, puis sur l'intégration de ces informations dans deux modèles de *boosting* performants, **LightGBM** et **CatBoost**.

Dans un premier temps, les données sont standardisées à l'aide de **StandardScaler** puis restructurées sous forme tridimensionnelle afin d'être compatibles avec l'architecture LSTM. Le réseau LSTM est composé d'une couche **LSTM de 64 unités** utilisant la fonction d'activation **tanh**, suivie d'une couche **Dense de 32 neurones** avec activation **ReLU** et d'une couche de sortie à un neurone. Le modèle est entraîné avec l'optimiseur **Adam** pendant **50 époques**, avec **batch_size = 64**, **validation_split = 0.2** et un mécanisme d'**Early Stopping** défini par **patience = 7** afin d'éviter le surapprentissage.

Les prédictions générées par le modèle LSTM sont ensuite utilisées comme une nouvelle caractéristique (LSTM *Feature*) ajoutée aux variables initiales du jeu de données. Ces nouvelles

données enrichies servent à entraîner un modèle **LightGBM** configuré avec **n_estimators = 300**, **learning_rate = 0.05** et **random_state = 42**, ainsi qu'un modèle **CatBoost** configuré avec **n_estimators = 300**, **learning_rate = 0.05**, **random_seed = 42** et **verbose = 0**.

La prédiction finale est obtenue par une stratégie d'agrégation de type **Ensemble Averaging**, où les prédictions de LightGBM et CatBoost sont combinées avec des poids égaux (**50 % – 50 %**). Cette approche permet de tirer profit à la fois de la capacité du LSTM à extraire des représentations complexes et de la puissance des algorithmes de *boosting* pour l'apprentissage sur des données tabulaires.

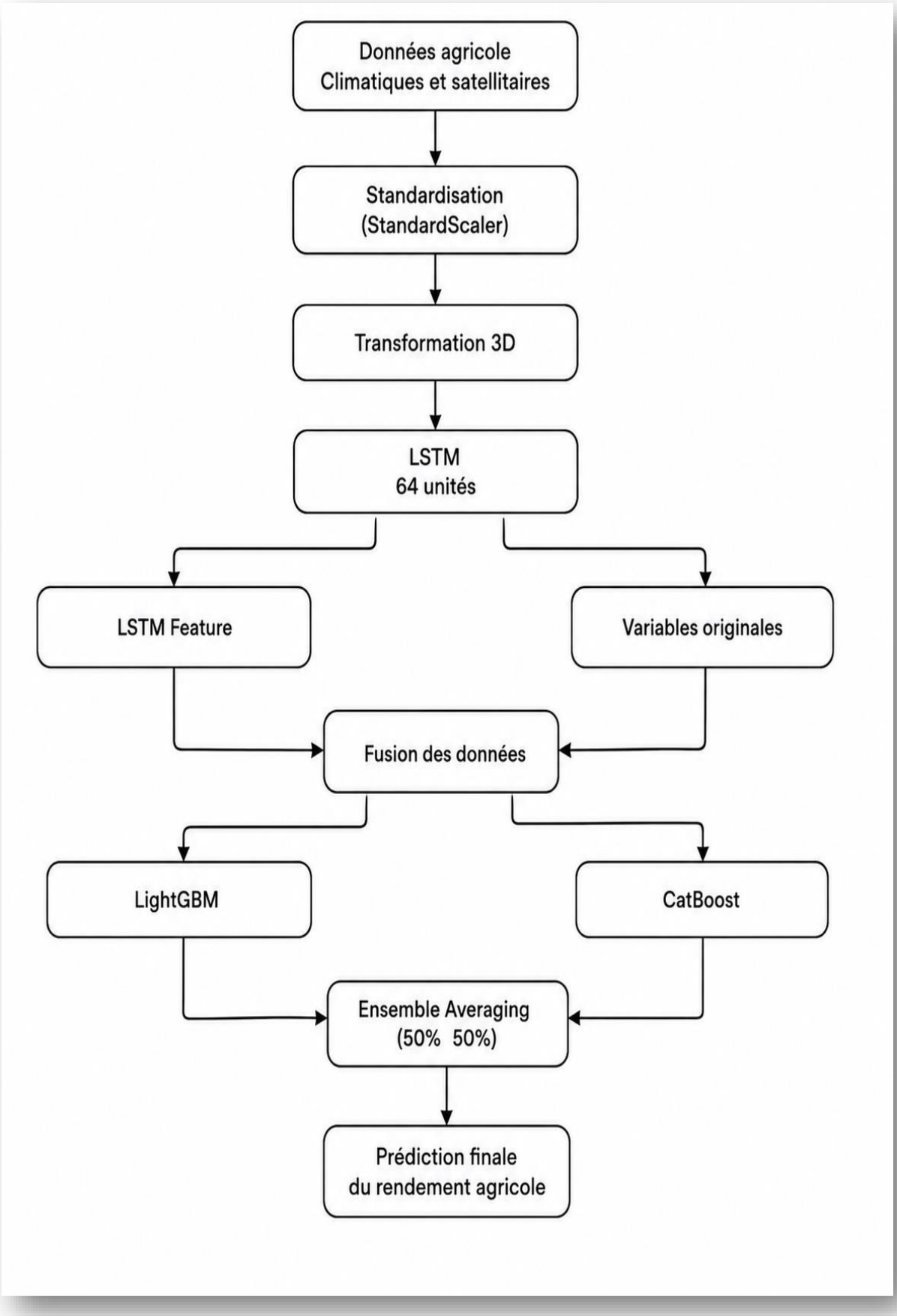


Figure 16: Architecture du modèle hybride proposé

3.13 Concluction

En résumé, ce chapitre a présenté de manière détaillée la méthodologie adoptée pour la mise en œuvre de notre système de prédiction du rendement agricole. Nous avons décrit le choix du jeu de données, les différentes étapes de prétraitement appliquées, ainsi que l'analyse exploratoire des données permettant de mieux comprendre leur structure et les relations entre les variables.

Nous avons également exposé les modèles d'apprentissage automatique et de Deep Learning utilisés dans cette étude, notamment Random Forest, SVR, LightGBM, CatBoost ainsi que les architectures neuronales telles que LSTM, GRU et MLP. Les stratégies de paramétrage des hyperparamètres ont été précisées afin d'optimiser les performances de chaque modèle et d'assurer une bonne capacité de généralisation.

Enfin, une approche hybride combinant LSTM et modèles de *boosting* a été présentée, ainsi qu'une analyse de l'importance des variables basée sur des méthodes d'interprétabilité telles que SHAP et Permutation Importance. Le chapitre suivant sera consacré à l'analyse comparative des résultats obtenus, en évaluant les performances de chaque modèle selon différents indicateurs afin d'identifier l'approche la plus efficace pour la prédiction du rendementa gricole.

Chapitre 4

Implémentation et expérimentation

4.1 Introduction

Ce chapitre constitue la phase de validation expérimentale de notre étude. Après avoir présenté les fondements théoriques de l'Intelligence Artificielle et des techniques de traitement de données dans les chapitres précédents, nous nous consacrons ici à l'analyse et à l'interprétation des résultats obtenus lors de l'entraînement et de l'évaluation des différents modèles prédictifs.

L'objectif principal est d'évaluer et de comparer la capacité de différentes architectures de pointe — allant des modèles de Machine Learning classiques (Gradient Boosting, XGBoost, CatBoost, LightGBM, SVR, Random Forest) aux réseaux de Deep Learning (LSTM, GRU, MLP) — à fournir des estimations précises du rendement céréalier sous diverses conditions climatiques en Algérie.

4.2 Environnement de développement

4.2.1 Langage de programmation

Python: est le langage principal utilisé dans ce travail. Il s'agit d'un langage de programmation interprété, simple et puissant, largement adopté dans les domaines de la science des données et de l'intelligence artificielle. Sa richesse en bibliothèques spécialisées en fait un choix optimal pour le développement rapide de modèles de Machine Learning et de Deep Learning [50].

4.2.2 Bibliothèques de Machine Learning

- **Scikit-learn** : Bibliothèque open-source utilisée pour l'implémentation des modèles classiques de Machine Learning tels que la régression, les forêts aléatoires et les machines à vecteurs de support. Elle a été choisie pour sa simplicité et sa robustesse [49].
- **XGBoost** : Bibliothèque optimisée et distribuée pour l'algorithme du Gradient Boosting. Elle est rapide, flexible et fréquemment utilisée dans les compétitions scientifiques et les projets industriels à grande échelle [51].
- **LightGBM** : Framework de Gradient Boosting basé sur les arbres de décision, conçu pour être distribué et efficace. Il offre un entraînement plus rapide, une utilisation mémoire réduite et unegestion native des variables catégorielles [52].

4.2.3 Bibliothèques de Deep Learning

- **TensorFlow** : Bibliothèque complète et open-source pour le calcul numérique utilisant des graphes de flux de données. Développée par Google, elle est principalement utilisée pour la construction et l'entraînement de réseaux de neurones profonds [53].
- **Keras** : API de haut niveau intégrée à TensorFlow, facilitant la conception rapide de modèles de Deep Learning, notamment les architectures récurrentes telles que les LSTM [54].

4.2.4 Traitement et visualisation des données

- **NumPy** : Bibliothèque fondamentale pour le calcul scientifique en Python. Elle fournit des objets array multidimensionnels (ndarray) et des fonctions mathématiques rapides pour les manipuler [55].
- **Pandas** : Bibliothèque de manipulation et d'analyse de données. Elle fournit les structures DataFrame et Series pour travailler de manière flexible et rapide avec des données tabulaires et temporelles [50].
- **Matplotlib** : Bibliothèque de tracé 2D permettant de créer des graphiques statiques, animés et interactifs. Largement utilisée pour visualiser les résultats des modèles [51].
- **Seaborn** : Bibliothèque construite sur Matplotlib, offrant une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs. Elle s'intègre nativement avec les structures de données Pandas [52].

4.2.5 Outils complémentaires

- **SHAP (SHapley Additive exPlanations)** : Méthode d'intelligence artificielle explicable (XAI) basée sur la théorie des jeux de Shapley, utilisée pour interpréter les prédictions des modèles et quantifier l'importance des variables d'entrée [56].
- **Joblib** : Ensemble d'outils permettant de sauvegarder et de charger efficacement de gros objets Python (tels que les modèles entraînés), tout en offrant une mise en cache sur disque pour les fonctions à exécution longue [57].
- **Streamlit** : Bibliothèque Python open-source pour créer rapidement des applications web interactives dédiées aux projets de data science et d'intelligence artificielle, sans nécessiter de connaissances en HTML, CSS ou JavaScript [58].
- **Google Colab** : Service cloud basé sur Jupyter Notebook permettant d'écrire et d'exécuter du code Python directement dans un navigateur, sans configuration locale. Il offre un accès gratuit à des ressources de calcul GPU et TPU, ce qui le rend particulièrement adapté aux tâches d'apprentissage automatique et de deep learning [59].
- **Visual Studio Code**: Visual Studio Code (VS Code) est un éditeur de code source léger et performant développé par Microsoft. Compatible avec Windows, macOS et Linux, il offre de nombreuses fonctionnalités facilitant le développement et prend en charge plusieurs langages de programmation grâce à son vaste écosystème d'extensions [60].

4.3 Évaluation des performances

Les expériences réalisées ont permis d'évaluer les performances de plusieurs modèles de Machine Learning et de Deep Learning appliqués à la prédiction du rendement agricole. Chaque modèle a été entraîné sur les données prétraitées issues des informations climatiques, satellitaires et agricoles, puis testé sur des données correspondant à une campagne agricole distincte afin

d'évaluer sa capacité de généralisation. Les résultats présentés dans les sections suivantes comparent les performances des différents modèles à l'aide de plusieurs métriques de régression couramment utilisées. Cette analyse permet d'identifier les approches les plus performantes et de déterminer le modèle le plus adapté à la prédiction du rendement agricole dans le contexte de cette étude.

4.3.1 Résultats des modèles ML

Le tableau suivant résume les métriques de performance obtenues sur l'ensemble de test pour chacun des six modèles de Machine Learning évalués.

Tableau 3 : Résultats de performance des modèles de Machine Learning

Modèle	MSE	RMSE	MAE	MAPE	MedAE	R^2	EVS	Adj. R^2
Gradient Boosting	14.05750	3.74933	3.150281	9.92620%	2.93735	0.77891	0.79578	0.77867
XGBoost	13.7464	3.70762	3.06883	9.43048%	2.83493	0.78380	0.795692	0.78357
CatBoost	13.3233	3.65012	3.08355	9.55061%	2.82723	0.790461	0.801731	0.79023
LightGBM	12.1659	3.48796	2.96908	9.07598%	2.71165	0.80866	0.81986	0.80841
SVR	20.88477	4.569986	3.87885	12.51761%	3.52030	0.46030	0.67611	0.67118
Random Forest	14.96128	3.86798	3.15759	9.64584%	2.89999	0.76470	0.790899	0.76444

Random Forest

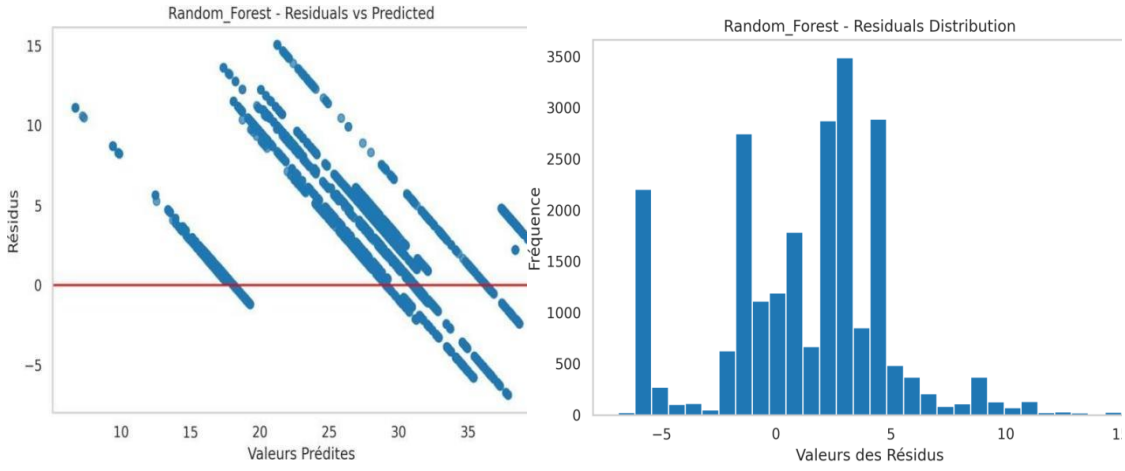


Figure 17 : Évaluation Random Forest

La figure 17 montre que le modèle Random Forest fournit des prédictions globalement précises. Les résidus sont répartis autour de zéro, ce qui traduit une bonne capacité de généralisation. Toutefois, la présence de structures linéaires dans le nuage des résidus et d'une distribution multimodale des erreurs révèle l'existence de différences de comportement selon les cultures ou les Wilayas, entraînant de légères surestimations ou sous-estimations du rendement

LightGBM

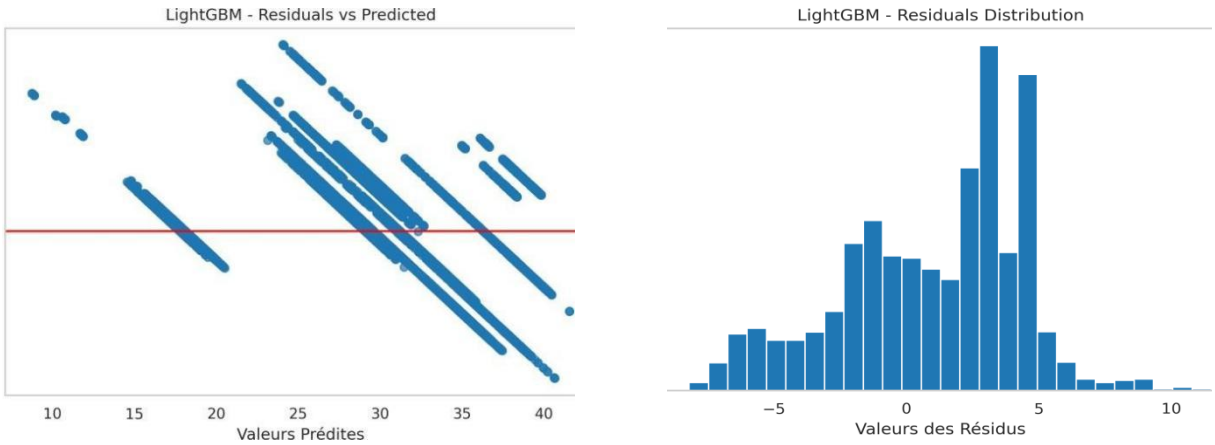


Figure 18 : Évaluation Light GBM

La figure 18 montre que le modèle LightGBM fournit des prédictions très précises avec une dispersion réduite des résidus autour de zéro. Le graphique met en évidence la persistance de structures linéaires associées à la nature des données agricoles et au décalage entre les observations de télédétection et les rendements officiels. L'histogramme des erreurs confirme également la bonne qualité des prédictions, malgré de légères variations de performance selon certaines Wilayas ou types de cultures.

XGBoost

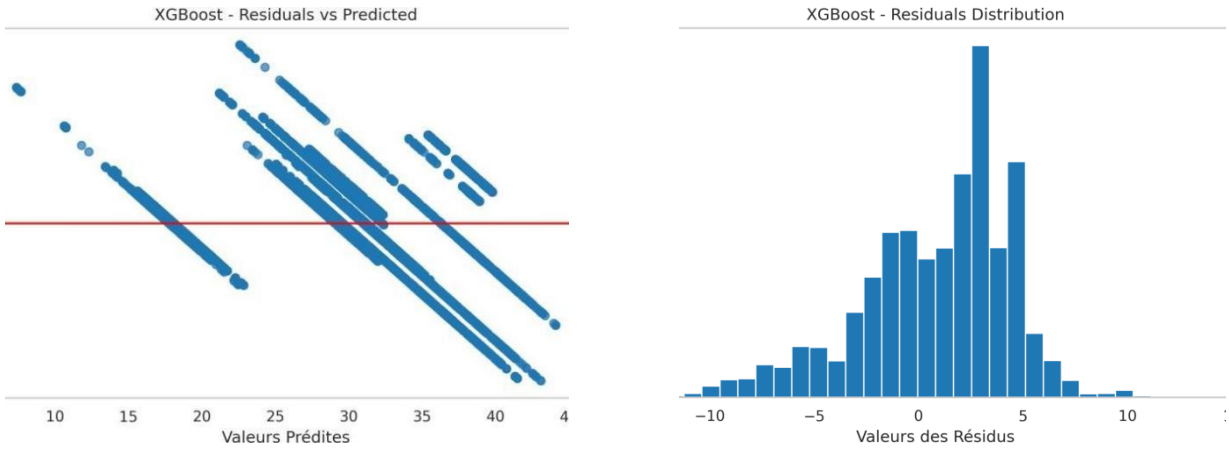


Figure 19 : Évaluation XGBoost

La figure 19 montre que le modèle XGBoost fournit des prédictions très précises avec des résidus majoritairement concentrés autour de zéro. Le graphique met en évidence des structures linéaires similaires à celles observées pour les autres modèles, reflétant le décalage entre les

données de télédétection et les rendements officiels. L'histogramme des erreurs confirme la robustesse du modèle, malgré de légères variations de performance entre certaines Wilayas et certains types de cultures.

Gradient Boosting

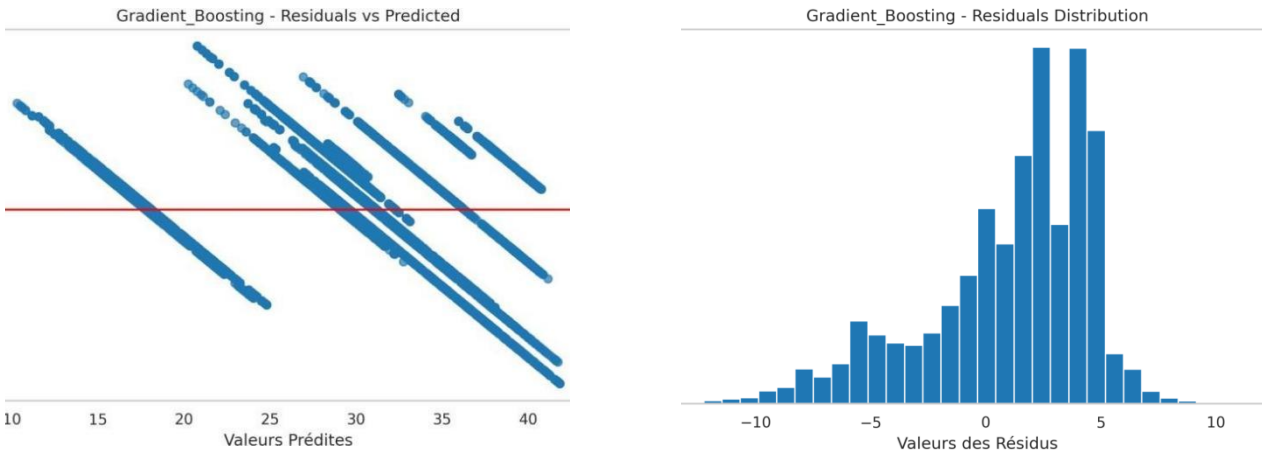


Figure 20 : Évaluation GBR

La figure 20 montre que le modèle Gradient Boosting (GBR) présente des performances globalement précises avec des résidus principalement concentrés autour de zéro. Le graphique des résidus en fonction des valeurs prédites révèle la persistance de structures linéaires inclinées similaires à celles observées avec les autres modèles, liées au décalage entre les données de télédétection continues et les rendements officiels fixes par région et par saison. L'histogramme des erreurs confirme une bonne convergence du modèle, malgré une légère multimodalité ndiquant des variations de performance selon certaines Wilayas et types de cultures.

CatBoost

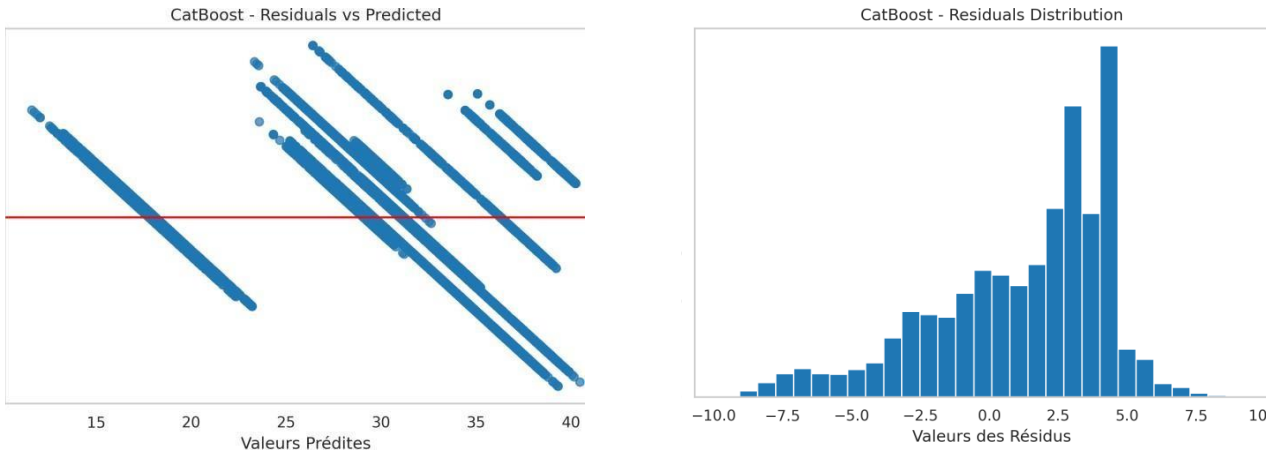


Figure 21: Évaluation CatBoost

La figure 21 montre que le modèle CatBoost offre des performances très précises avec des résidus fortement concentrés autour de zéro. Le graphique des résidus en fonction des valeurs prédites met en évidence la présence de bandes obliques parallèles similaires à celles observées avec les autres modèles de boosting, traduisant une erreur structurelle liée au décalage entre les données de télédétection continues et les rendements officiels fixes par région et par saison. L'histogramme des erreurs confirme une très bonne stabilité du modèle, malgré une légère multimodalité reflétant des variations selon certaines Wilayas et types de cultures.

SVR

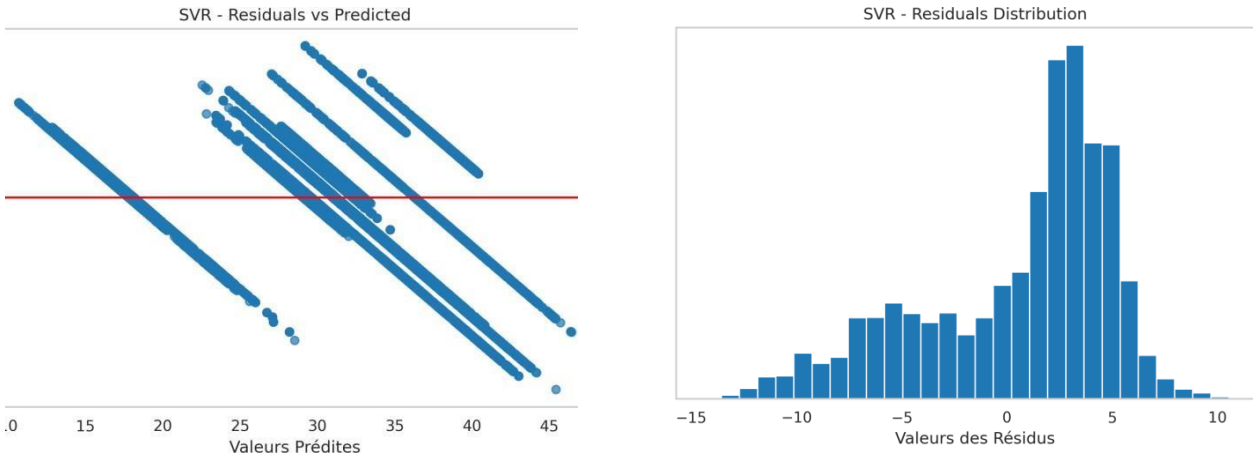


Figure 22 : Évaluation SVR

La figure 22 montre que le modèle Support Vector Regression (SVR) présente des performances globalement stables, avec des résidus majoritairement concentrés autour de zéro. Le graphique des résidus en fonction des valeurs prédites met en évidence la persistance des bandes obliques, mais avec une dispersion plus marquée par rapport aux modèles d'ensemble tels que LightGBM ou XGBoost, traduisant une sensibilité accrue au décalage entre les données de télédétection continues et les rendements officiels fixes par région et saison. L'histogramme des erreurs confirme une variance plus élevée et une légère multimodalité, reflétant des variations de performance selon les Wilayas et les types de cultures.

4.3.2 Résultats des modèles DL

Le tableau ci-dessous présente les métriques de performance obtenues pour les trois architectures de Deep Learning évaluées

Tableau 4: Résultats de performance des modèles de Deep Learning

Modèle	MSE	RMSE	MAE	MAPE	MedAE	R^2	EVS	Adj. R^2
LSTM	42.09086	6.48774	5.06083	19.72771%	4.07063	0.33803	0.48297	0.33731

MLP	30.13041	5.48911	4.46963	16.50031%	4.07044	0.526134	0.53531	0.52562
GRU	76.26704	8.73310	5.41866	22.24370%	4.15325	-0.19946	-0.10588	-0.20076

LSTM

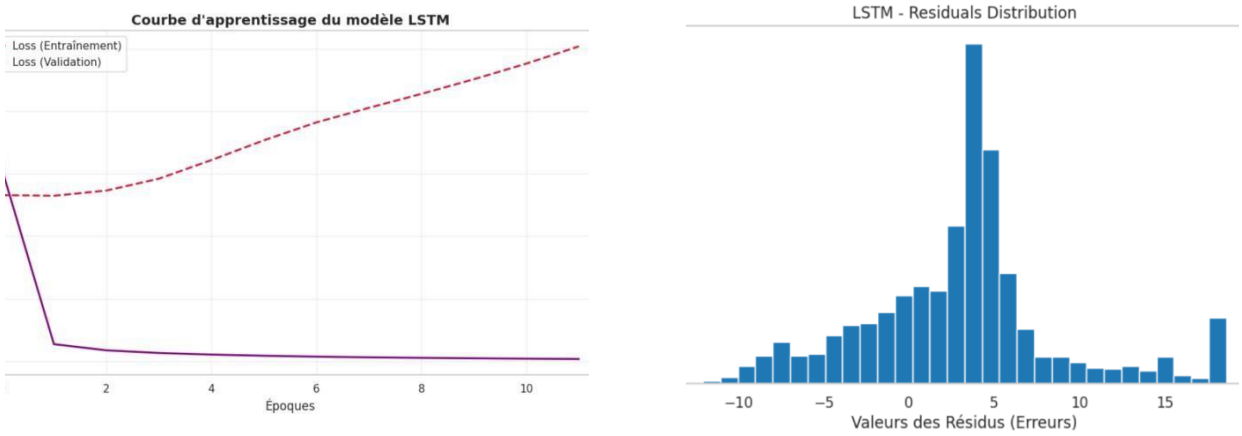


Figure 23 : Évaluation LSTM

L'analyse montre que le modèle présente des bandes obliques parallèles dans le graphique des résidus, reflétant une structure liée au décalage entre les prédictions continues issues de la télédétection et les rendements officiels (Yield_Official), qui restent fixes par région et par saison. Malgré cela, la distribution des erreurs reste fortement centrée autour de zéro, ce qui traduit une bonne précision globale et une faible variance des résidus.

MLP

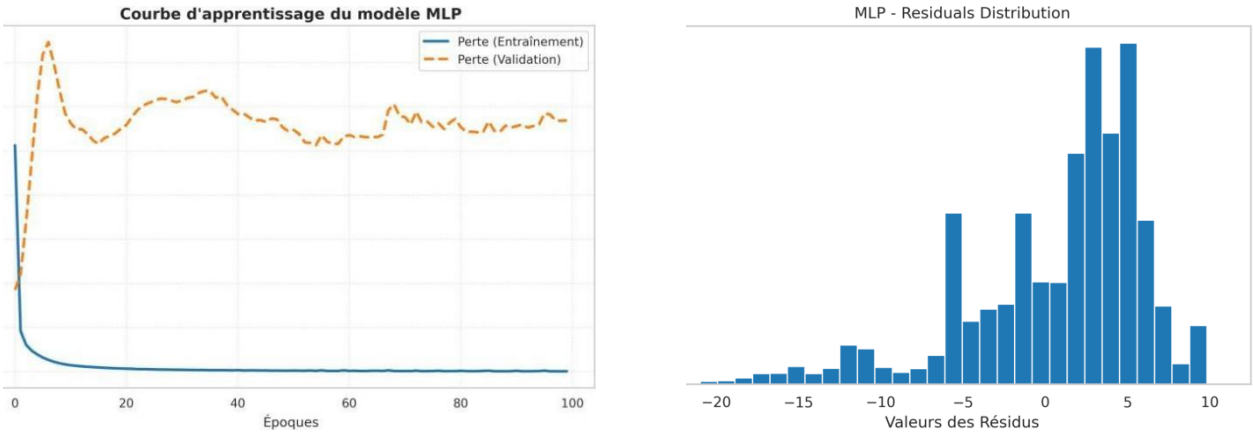


Figure 24 : Évaluation MLP

La courbe d'apprentissage du modèle MLP (Multi-Layer Perceptron) montre une diminution rapide et stable de la perte (MSE) dès les premières itérations, avec une convergence régulière des

courbes d'entraînement et de validation, sans écart significatif, ce qui indique une bonne stabilité et l'absence de surapprentissage. Par ailleurs, le graphique des résidus met en évidence des bandes obliques parallèles, reflétant un biais structurel lié au décalage entre les données satellitaires continues et les rendements officiels (Yield_Official) fixes par région et saison. L'histogramme confirme toutefois une forte concentration des erreurs autour de zéro, traduisant de bonnes performances globales du modèle.

GRU

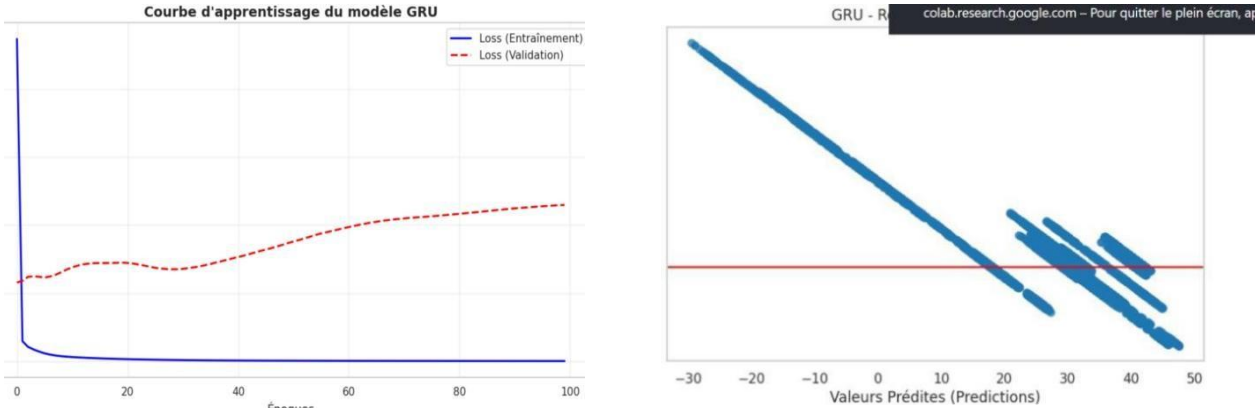


Figure 25 : Évaluation GRU

La courbe d'apprentissage du modèle GRU (Gated Recurrent Unit) montre une diminution rapide et stable de la perte (MSE) dès les premières époques, avec une convergence harmonieuse entre les courbes d'entraînement et de validation, sans écart significatif, ce qui confirme une bonne stabilité du modèle et l'absence de surapprentissage. Par ailleurs, l'analyse des résidus met en évidence des bandes obliques parallèles, traduisant un biais structurel lié au décalage entre les données de télédétection continues et les rendements officiels (Yield_Official) fixes par région et saison. L'histogramme des erreurs montre toutefois une forte concentration autour de zéro, indiquant de bonnes performances globales et une faible variance des prédictions.

4.3.3 Résultat de modèle hybride

Le tableau ci-dessous présente les métriques de performance obtenues pour le **modèle proposé hybride(LSTM +LightGBM+CatBoost).**

Tableau 5: Résultats de performance de modèle hybride

Modèle	MSE	RMSE	MAE	MAPE	MedAE	R^2	EVS	Adj. R^2
LGBM+CatBoost+LSTM	10.8887	3.29984	2.82660	8.58166%	2.80926	0.82874	0.84405	0.82855

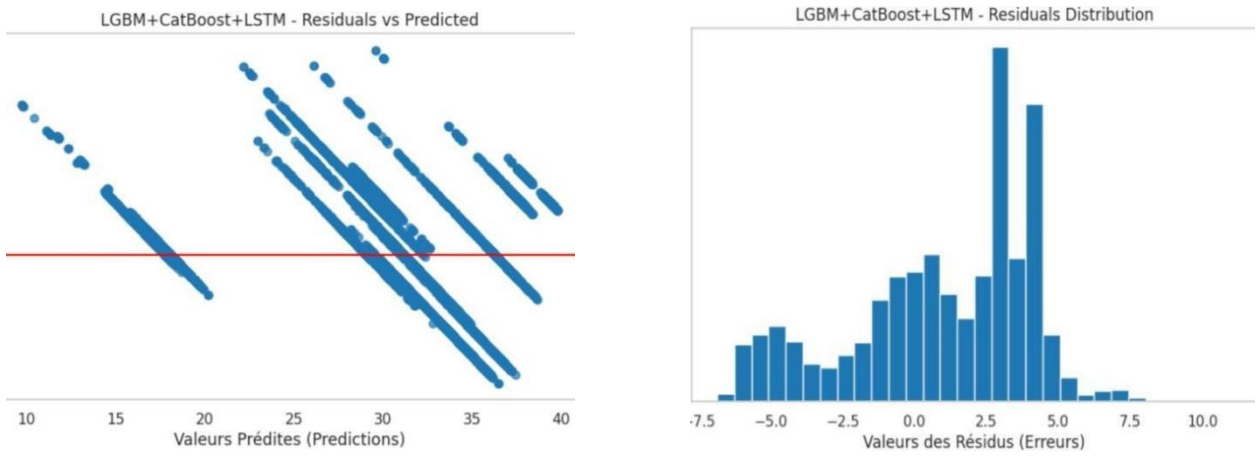


Figure 26 : Évaluation de notre modèle proposé hybride

La Figure 26 montre que le modèle hybride (LSTM + LightGBM + CatBoost) fournit des prédictions très précises avec des résidus fortement concentrés autour de zéro. Le graphique des résidus met en évidence la persistance de bandes obliques parallèles, liées au décalage entre les données de télédétection continues et les rendements officiels (Yield_Official) fixes par région et par saison. L'histogramme des erreurs confirme toutefois une faible dispersion et une excellente stabilité prédictive, traduisant l'efficacité de l'approche hybride dans l'amélioration de la précision des estimations du rendement agricole.

4.4 Comparaison des performances

Le tableau ci-dessous présente une synthèse comparative des meilleures performances obtenues par les modèles de Machine Learning et de Deep Learning, permettant une évaluation globale et équitable des différentes approches.

Tableau 6: Evaluation des meilleurs modèles ML et hybride

Modèle	Catégorie	MSE	RMSE	MAE	MAPE	R^2	Adj. R^2
LGBM+CatBoost+LSTM	Hybride	10.8887	3.29984	2.82660	8.58166%	0.82874	0.82855
LightGBM	ML	12.1659	3.48796	2.96908	9.07598%	0.80866	0.80841
CatBoost	ML	13.3233	3.65012	3.08355	9.55061%	0.790461	0.79023

4.4.1 Comparaison des performances selon les algorithmes

Les résultats obtenus montrent que certains modèles sont nettement plus performants que d'autres pour prédire les rendements agricoles. Les modèles de Deep Learning purs, notamment le **GRU**, affichent les résultats les plus limités avec un score R^2 négatif (-0.199), suivis par le **LSTM** ($R^2 = 0.338$) et le **MLP** ($R^2 = 0.526$). Ces performances modestes indiquent que

ces architectures, bien que puissantes, peinent à capturer les dynamiques de nos données, probablement en raison de la taille restreinte du jeu de données qui ne permet pas d'exploiter pleinement leur capacité d'apprentissage profond.

En revanche, les modèles de type *Boosting* (**LightGBM, CatBoost, XGBoost, Gradient Boosting**) offrent des résultats très satisfaisants. Parmi eux, le **LightGBM** se distingue comme l'un des plus efficaces avec un R^2 de **0.808**. Ils arrivent à bien capter la variabilité des rendements malgré la complexité des facteurs agricoles.

Cependant, le modèle le plus performant est notre modèle hybride **LGBM+CatBoost+LSTM**. Il surpasse tous les autres avec un R^2 de **0.828**, un RMSE minimal de **3.299** et une erreur absolue moyenne (MAE) de **2.826**. Ce modèle sera donc la référence retenue pour notre étude. Il offre le meilleur équilibre entre précision et robustesse, en combinant la puissance des algorithmes de *boosting* pour les données tabulaires et la gestion des dépendances temporelles via le LSTM.

Un point positif majeur est que ces modèles de *boosting* et le modèle hybride conservent une grande cohérence dans leurs métriques, ce qui confirme l'absence de surapprentissage. Si les modèles de Deep Learning isolés ont montré leurs limites dans ce contexte spécifique, les modèles combinés et les méthodes de *boosting* ont prouvé leur supériorité.

En conclusion, les approches basées sur le *boosting*, et plus particulièrement notre architecture hybride, sont les plus efficaces pour ce problème. Elles sont à la fois plus précises, mieux adaptées à la structure de nos données, et offrent une fiabilité accrue pour la prédiction des rendements agricoles dans notre zone d'étude.

Shape

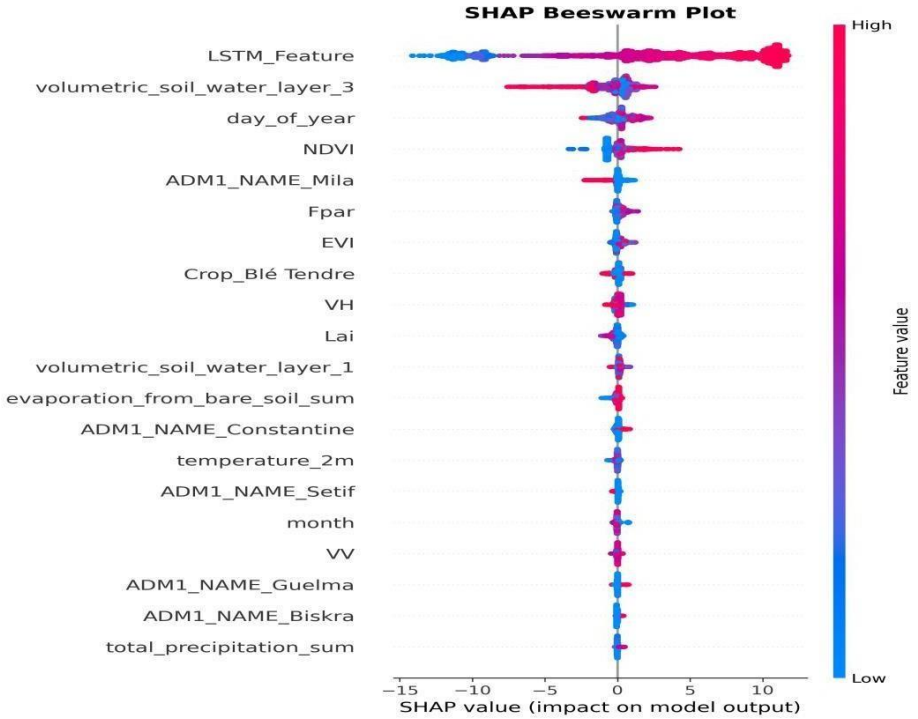


Figure 27: SHAP Beeswarm **et** Importance globale

Le diagramme « SHAP Beeswarm » illustre l'influence de chaque variable sur la prédiction du rendement des cultures. Chaque point représente une observation, sa position horizontale indiquant l'impact sur la prédiction (vers la droite pour une augmentation, vers la gauche pour une diminution), tandis que sa couleur (du bleu au rouge) reflète la valeur de la variable. Les trois caractéristiques les plus influentes sont LSTM_*Feature*, *volumetric_soil_water_layer*_3 et day_of_year. La LSTM_Feature exerce l'impact le plus fort avec une corrélation positive : des valeurs élevées (points rouges) augmentent significativement le rendement prédit. À l'inverse, *volumetric_soil_water_layer*_3 présente une relation inverse, où des valeurs élevées réduisent les prédictions, suggérant que l'excès d'humidité dans cette couche profonde est perçu négativement par le modèle. Enfin, *day_of_year* influence le modèle de manière plus nuancée, sans direction aussi marquée que les deux premières variables.

4.5 Application Web

4.5.1 Présentation Générale de l'Application

L'application web **AgriPredict**, développée avec **Streamlit** (Python) pour l'interface utilisateur et **FastAPI** comme moteur de calcul, représente une solution innovante d'aide à la décision destinée aux agriculteurs et aux décideurs agronomiques. Elle permet d'évaluer et de prédire le rendement agricole à partir de données satellitaires et climatiques, en s'appuyant sur un **modèle hybride d'apprentissage automatique** combinant les algorithmes **LSTM**, **CatBoost** et **LightGBM**, exploitant ainsi la complémentarité de ces approches pour maximiser la précision prédictive.

AgriPredict se distingue par plusieurs fonctionnalités majeures : la capacité d'effectuer des simulations prospectives jusqu'en 2030 pour la planification à long terme, une intelligence analytique régionale capable de diagnostiquer automatiquement les stress thermiques et hydriques spécifiques à chaque wilaya, ainsi qu'une interface multilingue (arabe, français, anglais) garantissant une accessibilité optimale. Grâce à une conception entièrement réactive, cette plateforme transforme des données géospatiales et climatiques complexes en rapports décisionnels simplifiés et actionnables.

4.5.2 Développement de l'Application

L'application repose sur une architecture **deux-tiers** : un frontend Streamlit communicant avec un backend FastAPI via des requêtes HTTP.

Backend

- **Langage :** Python
- **Architecture API :** FastAPI (utilisant Uvicorn) pour servir les prédictions de manière asynchrone

- Bibliothèques principales :
  - **Pandas et NumPy :** Pour la structuration des données d'entrée et le calcul des indices
  - **Scikit-learn :** Pour l'implémentation du modèle de régression du rendement
  - **SHAP (SHapley Additive exPlanations) :** Pour l'interprétation locale des prédictions et la visualisation de l'impact des variables
  - **Joblib :** Pour le chargement efficace du modèle pré-entraîné au format .pkl
  - **FPDF :** Pour la génération dynamique des certificats de rendement en PDF
  - **ArabicReshaper et Python-Bidi :** Pour le support correct du texte arabe dans les visualisations SHAP et les documents PDF générés

Frontend

- **Framework :** Streamlit, utilisé pour générer dynamiquement l'interface utilisateur en HTML, CSS et JavaScript
- **Composants UI :** Widgets natifs Streamlit (Sliders, Number Inputs, Tabs) pour une saisie intuitive des paramètres agronomiques (NDVI, EVI, Précipitations, etc.)
- Visualisation Graphique :
  - **Matplotlib :** Pour l'affichage des graphiques de barres SHAP illustrant la contribution de chaque variable
  - **Streamlit Line Chart :** Pour la visualisation de l'évolution temporelle du rendement prédit
  - **Gestion de l'état (State Management) :** Utilisation de st.session_state pour conserver les résultats de prédiction, les images SHAP et les données du rapport PDF pendant la session utilisateur
  - **Communication Frontend–Backend :** Bibliothèque requests pour l'envoi des données JSON du frontend Streamlit vers le backend FastAPI.

4.6 Développement du Modèle Final

Cette phase constitue le pont entre l'expérimentation et l'exploitation en production. Elle garantit que le modèle Hybride puisse fonctionner de manière autonome, reproductible et fiable au sein de l'application, indépendamment de l'environnement d'entraînement initial.

4.6.1 Sérialisation et Persistance avec Joblib

Après l'optimisation des hyperparamètres des modèles **CatBoost** et **LightGBM** ainsi que l'entraînement du réseau **LSTM**, chaque composant de notre architecture hybride est persisté sur disque. Les modèles basés sur les arbres de décision (**CatBoost** et **LightGBM**) sont sauvegardés

via la bibliothèque **Joblib**, tandis que le modèle **LSTM** est enregistré dans un format dédié aux réseaux de neurones. Ce choix technique, privilégié au format Pickle standard, repose sur deux

avantages critiques :

Efficacité mémoire et rapidité : Joblib est nativement optimisé pour la sérialisation des grands tableaux **NumPy** utilisés par CatBoost et LightGBM, ce qui réduit drastiquement la latence lors du chargement des modèles à chaque requête de prédiction.

Portabilité et robustesse en production : L'exportation de chaque composant sous forme de fichiers binaires isolés permet de figer les poids appris et de garantir une exécution stable en environnement de production, indépendamment de l'infrastructure ayant servi à la phase .

4.6.2 Pipeline de Prétraitement et Encodage des Variables

Afin de rendre les données catégorielles (telles que le nom de la Wilaya ou le type de culture) exploitables par les modèles du système hybride, un pipeline de prétraitement rigoureux a été intégré pour assurer la qualité et la pertinence des informations traitées :

Encodage catégoriel (One-Hot Encoding) : Les variables qualitatives sont transformées en vecteurs binaires. Cette approche méthodologique permet d'éviter d'imposer une relation d'ordre arbitraire entre les catégories, garantissant ainsi que le modèle traite chaque Wilaya et chaque culture comme des entités distinctes et indépendantes, assurant une équité statistique dans l'analyse.

Filtrage et validation de la cohérence physique : Avant toute phase d'inférence, une couche de validation est appliquée pour garantir la crédibilité des données entrantes. Ce mécanisme détecte et rejette systématiquement les valeurs aberrantes (ex. : indice NDVI > 1 ou température < −50 °C). Cette étape est cruciale, car elle empêche la transmission aux modèles de données dépourvues de sens physique, préservant ainsi la fiabilité agronomique des prédictions finales.

4.6.3 Ingénierie des fonctionnalités et Calcul du « Growth Factor »

Au-delà du traitement des données brutes, le système intègre une unité intelligente d'ingénierie des caractéristiques qui enrichit le vecteur d'entrée en temps réel, selon deux mécanismes complémentaires.

a) Estimation dynamique des variables cumulatives

b) Modélisation du cycle cultural – Growth Factor

Pour garantir la cohérence biologique des prévisions, une fonction logistique (sigmoïde) ajuste le rendement potentiel maximal fourni par le modèle hybride en fonction de l'avancement dans le cycle phénologique (d'octobre à juin). La formulation mathématique de cet ajustement est la suivante :

$$Y_final = Y_max \times [\, 1 / (1 + e\text{^}(-k \cdot (t - t_0))) \,]$$

Où :

- Y_final : rendement ajusté en temps réel (q/ha).

- Y_max : rendement potentiel maximal estimé par le modèle hybride.
- t : indice du mois courant dans le cycle cultural.
- k : pente de la courbe sigmoïde (vitesse de croissance).
- t_0 : point d'inflexion, correspondant à la phase de croissance maximale.

Cette approche garantit que les prévisions reflètent fidèlement la biomasse accumulée au moment de la consultation, évitant toute surestimation en début de saison.

4.7 Développement de l'API Backend

Le backend de l'application repose sur **FastAPI**, un framework web moderne et performant pour construire des APIs REST en Python. Il joue le rôle de moteur central, assurant la communication entre l'interface utilisateur et le modèle de prédiction.

4.7.1 Justification du Choix de FastAPI

FastAPI a été retenu pour ses caractéristiques techniques particulièrement adaptées aux exigences du système :

- **Performance :** fondé sur Starlette et Pydantic, il figure parmi les *frameworks* Python les plus rapides, comparable en performance aux frameworks Go et Node.js.
- **Asynchronisme (Async/Await) :** la gestion asynchrone des requêtes permet de traiter plusieurs appels simultanément sans bloquer le serveur, ce qui est essentiel pour les calculs intensifs comme l'inférence SHAP.
- **Validation automatique :** les classes Pydantic garantissent que chaque donnée envoyée par l'utilisateur est complète, correctement typée et conforme aux contraintes du modèle.
- **Documentation interactive automatique :** Swagger UI est généré automatiquement, permettant de tester chaque paramètre d'entrée de manière isolée avant l'intégration finale.

4.7.2 Cycle de Traitement du Point de Terminaison /predict

Lorsqu'une requête POST est adressée au point de terminaison /predict, le backend exécute un cycle de traitement structuré en cinq étapes successives :

1. **Prétraitement et encodage :** réception du fichier JSON du frontend, encodage des variables catégorielles (Wilaya, Culture) et activation de l'unité de *L'ingénierie des variables* pour le calcul des indices cumulatifs.
2. **Inférence hybride :** sollicitation du modèle hybride préalablement entraîné (LightGBM + CatBoost + LSTM) afin d'estimer le rendement brut (q/ha). La prédiction obtenue est ensuite ajustée à l'aide du **Growth Factor** pour tenir compte du stade phénologique actuel de la culture

et fournir une estimation plus réaliste du rendement final.

3. **Diagnostic régional :** comparaison des données climatiques saisies (température, humidité, indices de végétation) aux seuils critiques spécifiques à la zone agro-climatique de la Wilaya sélectionnée (Zone Saharienne, Hauts Plateaux ou Littoral).
4. **Génération de recommandations :** production dynamique de conseils agricoles ciblés en fonction des anomalies détectées (ex. : stress thermique si Tmax > 35 °C, ou déficit hydrique), tels que l'intensification de l'irrigation d'appoint ou l'apport de fertilisants azotés.
5. **Interprétation et réponse :** calcul des valeurs SHAP pour expliquer la décision du modèle et retour d'un objet JSON structuré contenant le rendement final, les graphiques d'impact et le rapport de diagnostic complet.

4.7.3 Sécurité et Documentation (Swagger UI)

FastAPI génère automatiquement une documentation interactive via **Swagger UI**, accessible depuis le navigateur. Cette fonctionnalité a permis de tester chaque paramètre d'entrée de manière isolée avant l'intégration finale, garantissant la robustesse et la cohérence des échanges de données entre le client et le serveur.

4.8 Développement de l'Interface Utilisateur (Frontend)

L'interface utilisateur constitue la couche d'interaction finale du système. Développée avec **Streamlit**, elle transforme un modèle d'intelligence artificielle complexe en un outil intuitif, accessible aux agriculteurs comme aux ingénieurs agronomes, sans nécessiter de compétences en développement web.

4.8. 1 Streamlit et Choix Technologiques

Streamlit a été sélectionné pour sa capacité à transformer des scripts Python en applications web interactives sans code HTML/CSS/JS. Pour la version, l'interface a été entièrement redessinée autour de quatre axes d'amélioration :

- **Réactivité et gestion de l'état (Session State) :** grâce à st.session_state, l'interface se met à jour dynamiquement sans rechargement de page, permettant à l'utilisateur de tester différents scénarios et de visualiser instantanément l'impact sur le rendement.
- **Composants UI avancés :** utilisation de st.sidebar pour les paramètres globaux, st.tabspour l'organisation thématique des données d'entrée, et Plotly pour les graphiques temporels interactifs.
- **Multilinguisme et traitement du texte arabe :** un dictionnaire de traduction intégré permet de basculer instantanément entre l'Arabe, le Français et l'Anglais. Les bibliothèques arabic_reshaperet python-bidi garantissent un affichage correct de l'arabe dans les graphiques SHAP et les rapports PDF.

- **IA explicable :** intégration des visualisations SHAP via Matplotlib pour interpréter les prédictions du modèle hybride et expliquer l'impact de chaque variable sur le rendement estimé, rendant ainsi l'intelligence artificielle accessible aux non-spécialistes.

4.8.2 Organisation Ergonomique de l'Interface

L'interface d'AgriPredict guide l'utilisateur à travers un flux de travail logique, de la saisie des paramètres jusqu'à l'obtention de rapports détaillés. Elle est structurée en trois zones fonctionnelles distinctes :

- Panneau de Configuration (Sidebar)

  Situé sur le côté gauche de l'écran, ce panneau permet de définir les paramètres contextuels de la prédiction : localisation géographique (sélection de la Wilaya), type de culture (Blé Dur, Blé Tendre, Avoine…) et période temporelle (mois et année).

- Formulaires Thématiques (Onglets)

  Afin d'éviter une surcharge cognitive, les variables d'entrée sont réparties en trois onglets interactifs permettant une saisie structurée par catégories :

  - **Onglet Climat :** température de l'air (2 m), évaporation, précipitations (mm).
  - **Onglet Télédétection :** indices de végétation (NDVI, EVI, LAI, Fpar).
  - **Onglet Humidité :** données radar (VV, VH) et humidité des couches superficielles (L1) et profondes (L3).

- Tableau de Bord des Résultats

  Une fois l'inférence déclenchée, les résultats sont présentés sous trois angles complémentaires :

- **Analyse Diagnostique :** détection des anomalies (stress thermique, déficit hydrique) et proposition de solutions concrètes pour optimiser le rendement.
- **Évolution Temporelle :** courbe de croissance prévisionnelle (sigmoïde) affichant la trajectoire du rendement du semis à la récolte.
- **Interprétation SHAP :** module d'IA explicable identifiant et classant les facteurs ayant eu l'impact le plussignificatif sur le résultat final.

4.8.3 Visualisations Avancées et Exportation

Au-delà de la simple prédiction numérique, le frontend intègre des outils d'aide à la décision permettant une interprétation approfondie et une traçabilité complète des résultats :

- **Graphiques SHAP :** La bibliothèque Matplotlib est utilisée pour générer des visualisations SHAP (SHapley Additive exPlanations), illustrant la contribution positive ou négative de chaque variable d'entrée sur le rendement final prédit. Cette approche s'inscrit dans les principes de l'IA Explicable (Explainable AI), permettant d'interpréter les prédictions du modèle hybride

(LightGBM + CatBoost + LSTM) et d'identifier l'impact de chaque variable sur le rendement estimé, y compris pour les utilisateurs non spécialistes.

- **Génération de Rapport PDF :** La bibliothèque **FPDF** est exploitée pour convertir l'ensemble des résultats de la session en un **certificat de rendement officiel téléchargeable**. Ce document structuré regroupe l'identité de l'agriculteur, les paramètres agro-climatiques saisis, le rendement prédit, le diagnostic régional et les recommandations générées, offrant ainsi un support de communication fiable entre les parties prenantes.

4.8.4 Présentation des Interfaces et Fonctionnalités

- **Diagramme de Flux de l'Application**

La figure 28 présente le diagramme de flux complet de l'application **AgriPredict** , illustrant le cheminement des données depuis la saisie utilisateur jusqu'à la génération des rapports finaux.

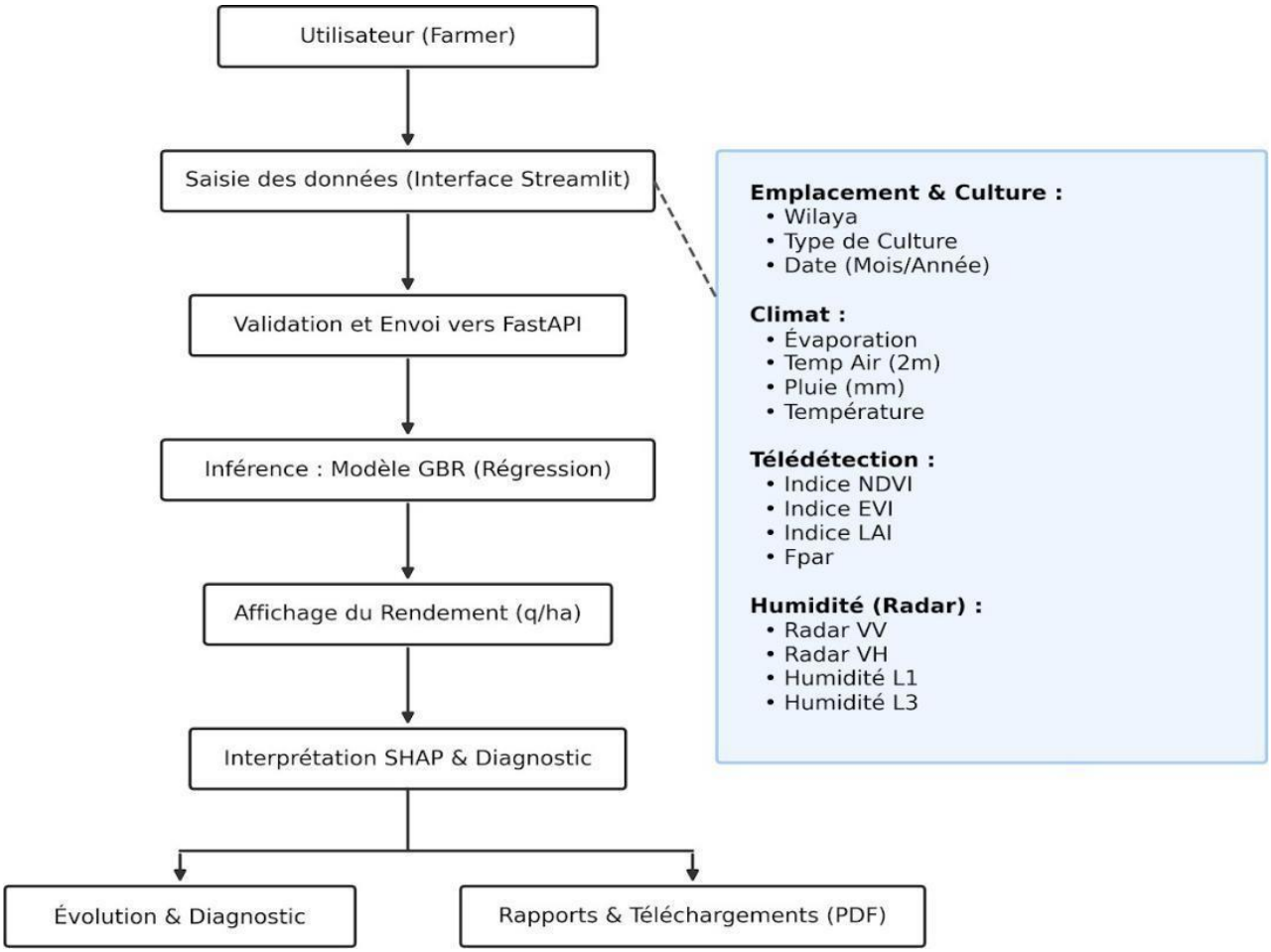


Figure 28 : Diagramme de flux détaillé de l'application **AgriPredict**

- **Page d'accueil**

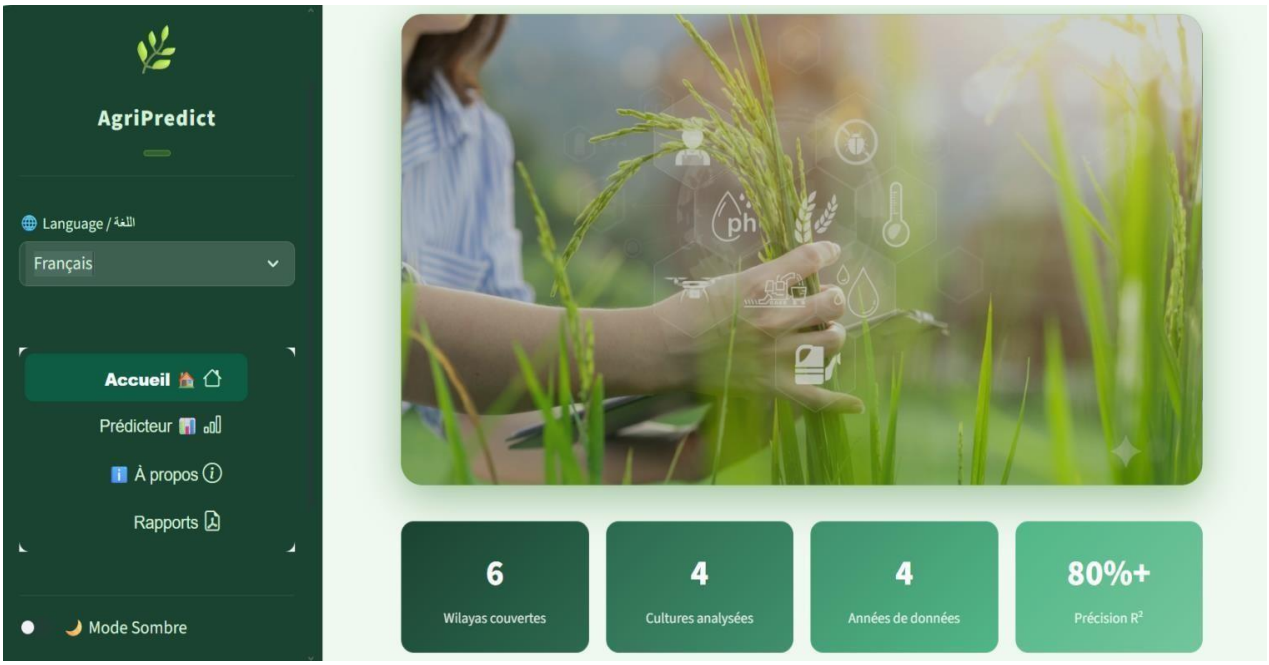
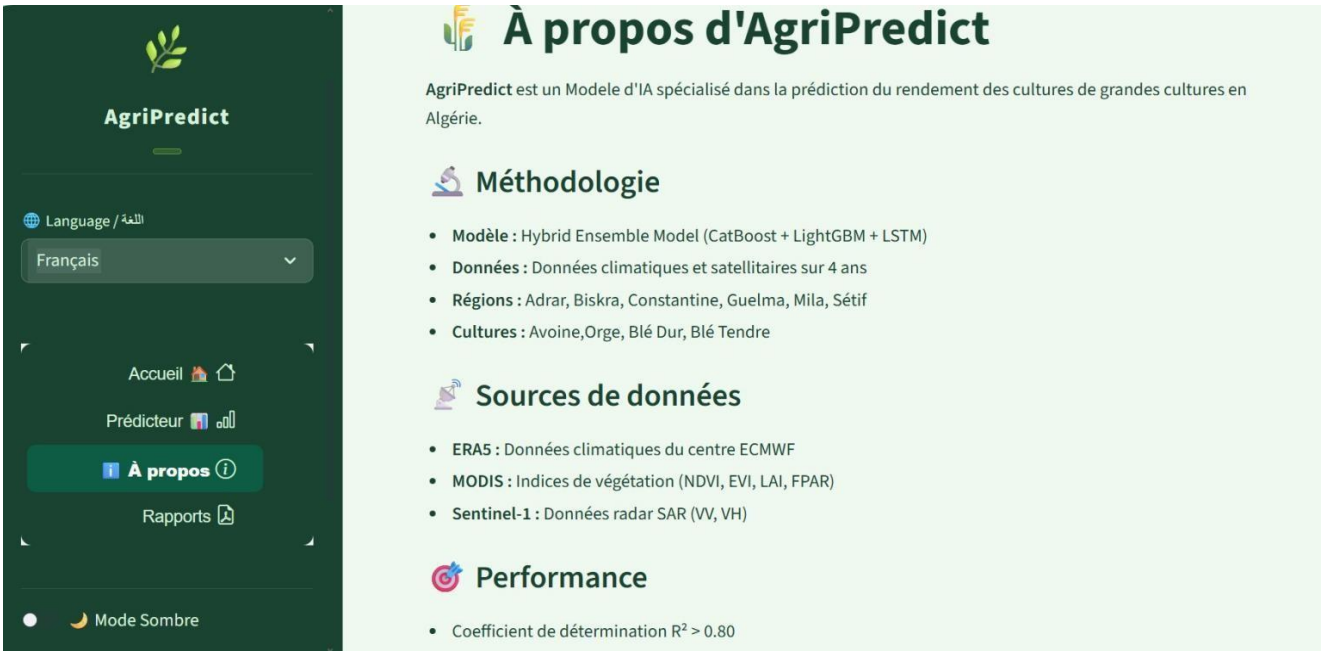


Figure 29 : L'interface d'accueil **AgriPredict**

La page d'accueil présente l'application **AgriPredict**, un véritable outil d'aide à la décision conçu pour l'évaluation et la prédiction des rendements agricoles. Grâce à une interface intuitive, elle permet à l'utilisateur d'accéder facilement à la plateforme de prédiction via un bouton dédié.

- **Page À propos du Modèle**



Cette page présente une vue d'ensemble d'AgriPredict et de sa méthodologie : un modèle hybride

Figure 30 : Interface À propos du Modèle

(CatBoost + LightGBM + LSTM) entraîné sur 4 ans de données climatiques et satellitaires, couvrant plusieurs Wilayas quatre cultures (Avoine, Orge, Blé Dur, Blé Tendre). Elle mentionne également les sources de données ERA5, MODIS, Sentinel-1) ainsi que l'indicateur de performance global du modèle.

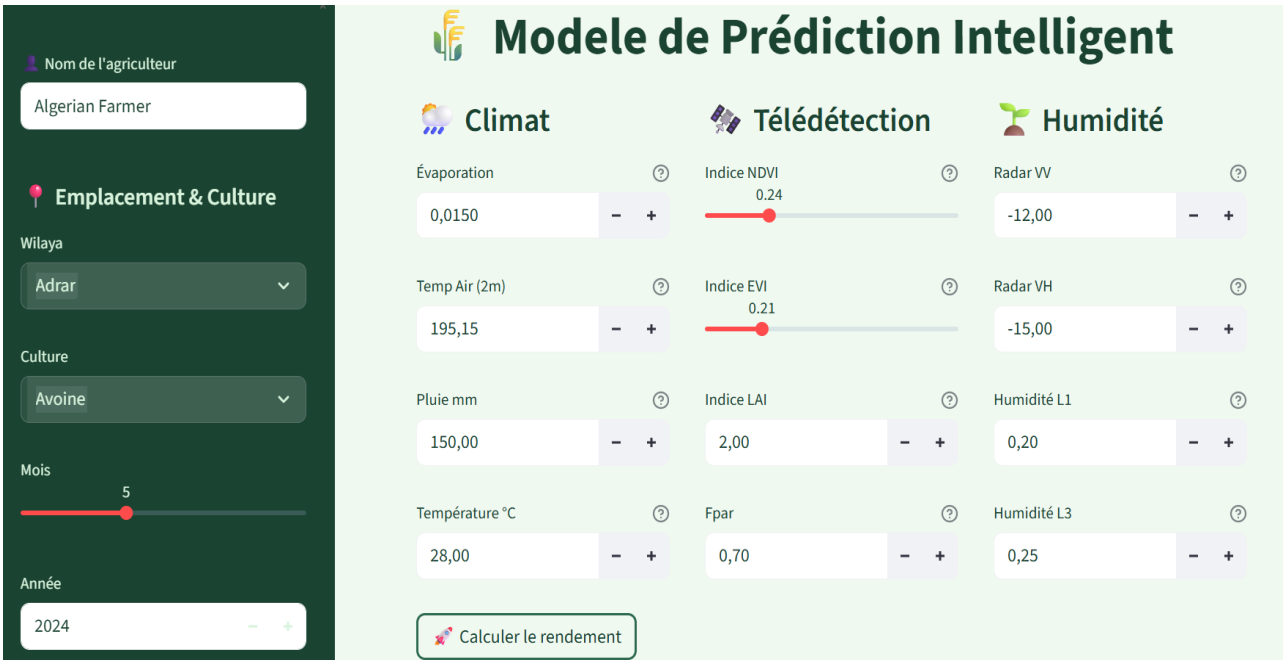


Figure 31 : Interface de saisie des données et système d'aide interactive (Infobulles).

La figure 31 illustre l'interface principale de saisie. L'utilisateur y renseigne les paramètres organisés en trois onglets (Climat, Télédétection, Humidité du sol). Deux fonctionnalités d'assistance avancées renforcent la fiabilité des données saisies :

- Infobulles contextuelles : chaque champ de saisie est accompagné d'une bulle d'aide interactive s'affichant au survol, précisant les unités requises (mm, °C) et l'importance agronomique de la variable (ex. : rôle des indices NDVI/LAI dans l'estimation de la biomasse).
- Traitement réactif : dès l'activation du bouton « Calculer le rendement », le système affiche instantanément le rendement estimé (q/ha), le diagnostic régional automatisé et les recommandations ciblées, sur la même page.

- **Module d'Analyse de Corrélation et Alertes Intelligentes**

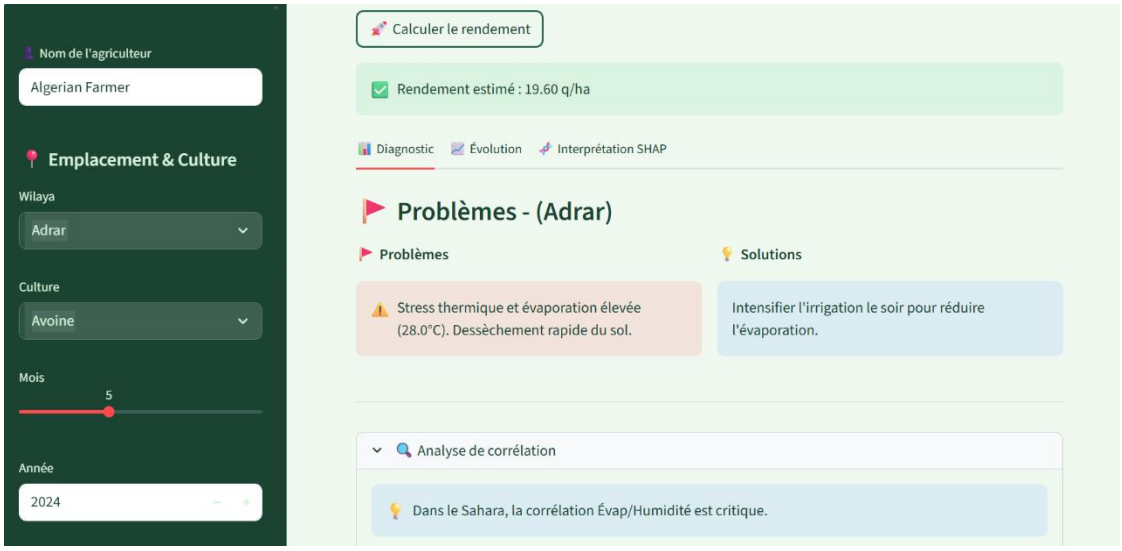


Figure 32 : Module d'analyse de corrélation et alertes intelligentes basées sur les spécificités climatiques.

Comme illustré en figure 32 un module d'intelligence métier génère des conseils dynamiques en identifiant les relations critiques entre variables agro-climatiques. Par exemple, dans les zones sahariennes, une hausse brutale de température entraîne une chute exponentielle de l'humidité du sol, et le système génère alors des alertes préventives pour anticiper les baisses de rendement liées à ces chocs thermiques.

- **Suivi de l'Évolution Temporelle du Rendement**



Figure 33 : Courbe de croissance prévisionnelle basée sur le modèle sigmoïde (Octobre - Juin).

L'onglet « Évolution » (figure 33) présente une courbe sigmoïde illustrant la progression du rendement tout au long du cycle cultural (octobre à juin). La prédiction actuelle, marquée par une étoile rouge, est ainsi positionnée par rapport au potentiel maximal de fin de saison, offrant une lecture

intuitive de l'avancement de la culture.

- **Interprétation Globale et Explicabilité via shap**

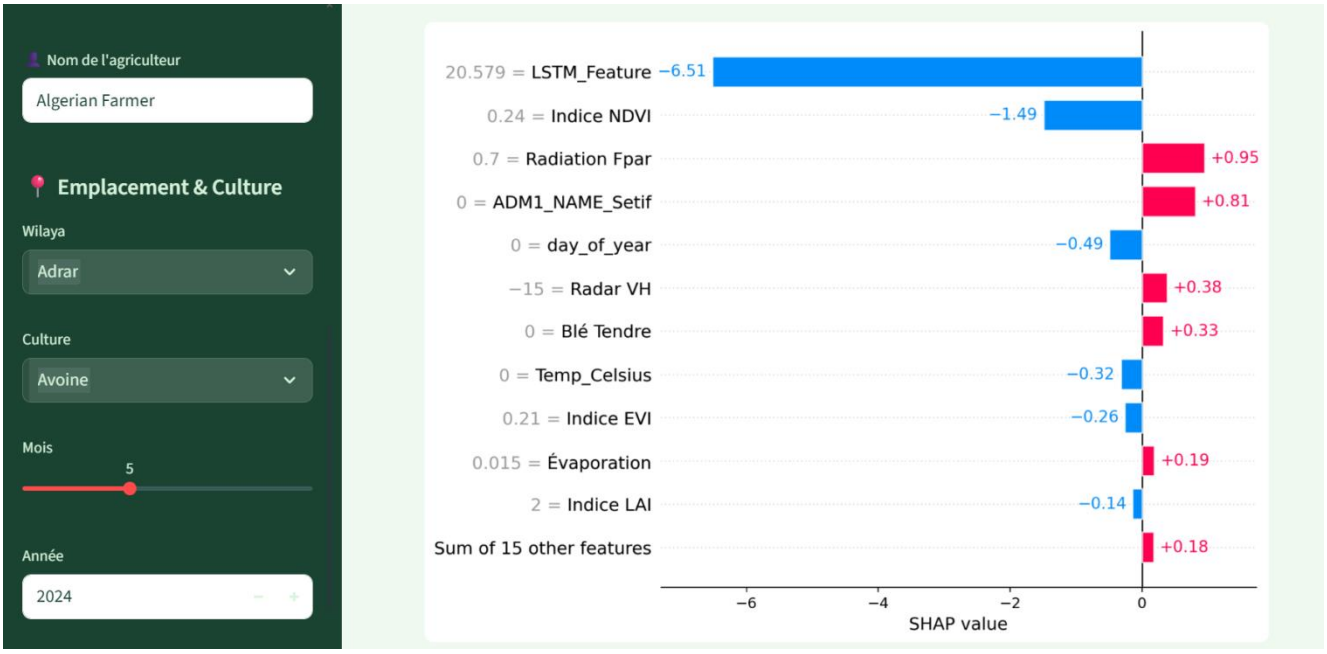
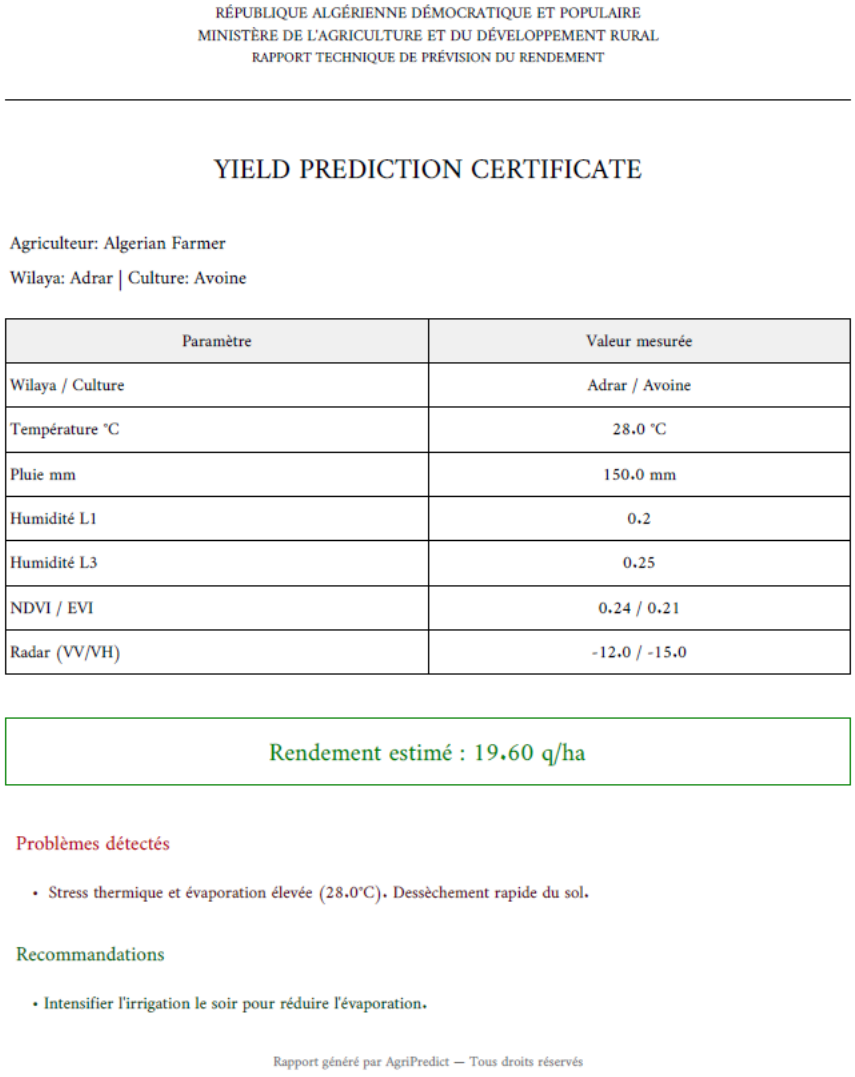


Figure 34 : Analyse de l'importance des variables via les valeurs SHAP.

L'onglet « Interprétation SHAP » (34) montre l'influence de chaque variable sur la prédiction : les barres rouges représentent les facteurs favorables au rendement (ex. : précipitations cumulées élevées, EVI favorable), tandis que les barres bleues indiquent les facteurs pénalisants (ex. : évapotranspiration excessive, faible Fpar). Cette transparence constitue le cœur de l'IA Explicable (XAI) intégrée au système.

- **Génération du Rapport Technique (Certificat PDF)**



RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'AGRICULTURE ET DU DÉVELOPPEMENT RURAL
RAPPORT TECHNIQUE DE PRÉVISION DU RENDEMENT

YIELD PREDICTION CERTIFICATE

Agriculteur: Algerian Farmer

Wilaya: Adrar | Culture: Avoine

Paramètre	Valeur mesurée
Wilaya / Culture	Adrar / Avoine
Température °C	28.0 °C
Pluie mm	150.0 mm
Humidité L1	0.2
Humidité L3	0.25
NDVI / EVI	0.24 / 0.21
Radar (VV/VH)	-12.0 / -15.0

Rendement estimé : 19.60 q/ha

Problèmes détectés

- Stress thermique et évaporation élevée (28.0°C). Dessèchement rapide du sol.

Recommandations

- Intensifier l'irrigation le soir pour réduire l'évaporation.

Rapport généré par AgriPredict — Tous droits réservés

Figure 35 : Certificat technique de prévision du rendement généré au format PDF

L'application permet d'exporter un Yield Prediction Certificate en PDF (figure 35), récapitulant l'identité de l'agriculteur, les paramètres agro-climatiques saisis et le rendement prédit (q/ha). Ce rapport sert de support de décision fiable pour les organismes agricoles, assurances et coopératives.

4.10 Conclusion

Ce chapitre a présenté la validation expérimentale de notre étude. Après une comparaison de plusieurs modèles de Machine Learning et de Deep Learning, le modèle hybride LSTM+CatBoost+LightGBM s'est révélé le plus performant ($R^2 = 0.828$, RMSE = 3.299), confirmant l'intérêt de combiner boosting et apprentissage profond. L'analyse SHAP a renforcé l'interprétabilité des résultats en identifiant les variables les plus influentes.

Nous avons ensuite présenté l'application AgriPredict, basée sur une architecture Streamlit/FastAPI, en détaillant son pipeline de traitement, son moteur d'inférence hybride, son

interface utilisateur et ses fonctionnalités (diagnostic régional, suivi du rendement, SHAP, certificat PDF).

Enfin, les limites de l'étude (résolution des données, facteurs agronomiques manquants, portée géographique) ont été soulignées, ouvrant des perspectives pour des travaux futurs.

Conclusion générale

Au terme de ce mémoire de master, il apparaît clairement que l'intelligence artificielle représente aujourd'hui un levier majeur pour améliorer la prédiction du rendement agricole dans le contexte de l'agriculture intelligente. Face aux défis croissants liés aux changements climatiques, à la pression sur les ressources naturelles et à l'augmentation des besoins alimentaires, le recours à des méthodes avancées d'analyse et de prédiction devient indispensable pour accompagner une gestion agricole plus efficace et durable.

L'objectif principal de ce travail était de développer et d'évaluer des modèles capables d'estimer avec précision le rendement agricole à partir de données issues de différentes sources. Pour atteindre cet objectif, une démarche complète a été mise en œuvre, comprenant la collecte et le prétraitement des données, l'analyse exploratoire, la sélection des variables pertinentes, la conception de plusieurs modèles de Machine Learning et de Deep Learning, ainsi que l'évaluation comparative de leurs performances.

Les expérimentations réalisées ont permis de démontrer l'intérêt de l'utilisation conjointe des données climatiques, agricoles et satellitaires pour améliorer la qualité des prédictions. Les résultats obtenus montrent que les modèles basés sur les techniques de boosting, notamment LightGBM et CatBoost, offrent des performances particulièrement élevées grâce à leur capacité à modéliser efficacement les relations non linéaires présentes dans les données. Les modèles de Deep Learning, tels que les réseaux LSTM, ont également montré leur pertinence pour exploiter la dimension temporelle des informations agricoles.

Une contribution de ce travail réside dans la proposition d'un nouvelle **modèle hybride**, combinant les forces complémentaires du Deep Learning et du Machine Learning avancé. En particulier, l'architecture intégrant LSTM pour la modélisation des dépendances temporelles et les algorithmes de boosting LightGBM et CatBoost pour l'apprentissage des relations non linéaires s'est révélée la plus performante parmi les modèles étudiés. Ce résultat confirme l'intérêt des approches hybrides pour améliorer la précision et la robustesse des systèmes de prédiction du rendement agricole.

Par ailleurs, l'intégration des techniques d'intelligence artificielle explicable à travers la méthode SHAP a permis d'améliorer l'interprétabilité des résultats. Cette approche offre une meilleure compréhension de l'influence des différentes variables sur les prédictions et renforce la confiance des utilisateurs dans les décisions produites par le système.

Au-delà de l'aspect expérimental, ce travail a abouti au développement d'une application intelligente d'aide à la décision basée sur Streamlit et FastAPI. Cette plateforme permet non seulement d'estimer le rendement agricole, mais également de visualiser les résultats, d'analyser l'importance des variables et de générer des rapports détaillés destinés aux agriculteurs et aux

décideurs. Cette réalisation constitue une étape importante vers l'exploitation concrète des technologies d'intelligence artificielle au service de l'agriculture de précision.

Enfin, les résultats obtenus confirment que l'intégration du Machine Learning, du Deep Learning, des données satellitaires et des techniques d'intelligence artificielle explicable constitue une solution prometteuse pour améliorer la prédiction du rendement agricole. Ce travail contribue ainsi au développement de systèmes intelligents capables de soutenir les acteurs du secteur agricole dans leurs décisions stratégiques et opérationnelles, tout en favorisant une agriculture plus performante, durable et résiliente face aux défis futurs.

Défis du travail

Malgré Les bons résultats obtenues par nos modèles, cette étude présente certaines limites qu'il convient de souligner afin de situer le cadre d'application de nos résultats :

- **Disponibilité et résolution des données :** L'une des principales limites réside dans la résolution spatiale et temporelle des données satellitaires et climatiques utilisées. Une résolution plus fine pourrait capturer des variations micro-climatiques locales qui influencent la productivité parcelle par parcelle
- **Facteurs agronomiques non inclus :** Nos modèles se basent essentiellement sur la télédétection (NDVI) et les variables climatiques. Cependant, d'autres facteurs agronomiques déterminants — tels que les pratiques culturales (dates de semis, types et doses d'engrais), les attaques de ravageurs ou les maladies phytosanitaires — n'ont pas pu être pleinement **intégrés** dans nos modèles, faute de disponibilité de données numériques structurées et géoréférencées sur ces variables à l'échelle nationale. Or, ces facteurs peuvent avoir un impact significatif sur le rendement final, indépendamment des conditions climatiques ou de l'état de la végétation observé par télédétection.
- **Dépendance aux données historiques :** Les modèles de Machine Learning et de Deep Lear- ning (comme le LSTM) s'appuient sur des cycles passés. En cas de changements climatiques extrêmes et inédits (événements "hors norme"), les modèles pourraient rencontrer des difficultés à maintenir le même niveau de précision
- **Complexité des modèles combinés :** Bien que les modèles de type Boosting (Gradient Boos- ting, XGBoost) soient très précis, leur interprétabilité reste parfois complexe par rapport à des modèles statistiques linéaires simples, ce qui nécessite des outils supplémentaires pour expli- quer les décisions aux décideurs non techniciens.
- **Portée géographique :** Bien que l'étude se concentre sur l'Algérie, les spécificités pédologiques (nature du sol) varient énormément d'une région à une autre, et l'absence de cartes de sols détaillées dans le dataset limite la capacité du modèle à intégrer la fertilité intrinsèque des terres.

En conclusion, ces limites ne remettent pas en cause la validité globale des résultats obtenus.

Elles ouvrent néanmoins des perspectives prometteuses pour des recherches futures, notamment l'intégration de données multisources à plus haute résolution spatiale et temporelle, ainsi que le recours à des techniques d'intelligence artificielle explicable (XAI) permettant une meilleure interprétabilité des modèles auprès des décideurs.

Perspectives

Malgré les résultats encourageants obtenus, plusieurs pistes d'amélioration peuvent être envisagées pour enrichir ce travail :

- Intégrer des données supplémentaires relatives aux sols, aux pratiques culturales et aux conditions agronomiques afin d'améliorer davantage la précision des prédictions.
- Exploiter des images satellitaires à plus haute résolution pour réaliser des estimations plus fines à l'échelle de la parcelle agricole.
- Développer une version mobile du système permettant aux agriculteurs d'accéder aux prédictions et recommandations directement sur le terrain.
- Intégrer des capteurs IoT pour collecter des données en temps réel et renforcer les capacités de suivi et de prévision.
- Étendre l'approche à d'autres cultures et à différentes régions afin d'évaluer la robustesse et la capacité de généralisation du système.

Ces perspectives ouvrent la voie à la conception de plateformes agricoles intelligentes toujours plus performantes, capables d'accompagner efficacement la transition vers une agriculture de précision.

Bibliographie

[1] FAO. (2021). Making agrifood systems more resilient to shocks and stresses: The State of Food and Agriculture 2021. Food and Agriculture Organization of the United Nations.

[2] Pörtner, H. O., et al. (2022). Climate change 2022: Impacts, adaptation and vulnerability. IPCC. Cambridge University Press.

[3] Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big data in smart farming – A review. *Agricultural Systems*, 153, 69–80.

[4] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90.

[5] Zhang, C., & Kovacs, J. M. (2012). The application of small unmanned aerial systems for precision agriculture: A review. *Precision Agriculture*, 13(6), 693–712.

[6] You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. *AAAI*, 31(1), 4559–4566.

[7] Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11), 1443–1452.

[8] Alexandratos, N., & Bruinsma, J. (2012). World agriculture towards 2030/2050: The 2012 revision (ESA Working Paper No. 12-03). FAO.

[9] Ali, Z., Muhammad, A., Lee, N., Waqar, M., & Lee, S. W. (2025). Artificial intelligence for sustainable agriculture: A comprehensive review of AI-driven technologies in crop production. *Sustainability*, 17(5), 2281.

[10] Singh, A. K., Junior, F. N. F., Mainsah, N. L., & Abdoul-Rahmane, B. (2024). Enabling data collection and analysis for precision agriculture in smart farms. *IEEE Transactions on AgriFood Electronics*, 3(1), 69–85.

[11] Jabed, M. A., & Murad, M. A. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches. *Heliyon*, 10(24), e40891.

[12] van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709.

[13] Pugh, N. A., et al. (2024). Yield prediction in a peanut breeding program using remote sensing data and machine learning algorithms. *Frontiers in Plant Science*, 15, 1339864.

[14] Betew, A. G., et al. (2025). Historical and contemporary crop yield prediction models: Key lessons and innovations. *Smart Agricultural Technology*, 10, 101672.

[15] Bouregaa, T. (2026). Assessment of machine learning models for predicting cereal production in Algeria based on climatic and agricultural variables. *Modeling Earth Systems and Environment*, 12(1), 27.

[16] Ishaq, R. A. F., et al. (2025). Geospatial robust wheat yield prediction using machine learning and

integrated crop growth model and time-series satellite data. *Remote Sensing*, 17(7), 1140.

[17] Schowengerdt, R. A. (2006). *Remote sensing: Models and methods for image processing* (3rd ed.). Elsevier.

[18] Jensen, J. R. (1996). *Introductory digital image processing: A remote sensing perspective* (2nd ed.). Prentice Hall.

[19] Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, 236, 111402.

[20] Joshi, A., Pradhan, B., Gite, S., & Chakraborty, S. (2023). Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: A systematic review. *Remote Sensing*, 15(8), 2014.

[21] Adriko, K., et al. (2025). From MODIS to Sentinel-2: A regional comparative analysis of crop-yield prediction with matched spatiotemporal data. *IEEE JSTARS*, 18, 1–18.

[22] Wulder, M. A., et al. (2022). Fifty years of Landsat science and impacts. *Remote Sensing of Environment*, 280, 113195.

[23] Nasirzadehdizaji, R., et al. (2019). Sensitivity analysis of multi-temporal Sentinel-1 SAR parameters to crop height and canopy coverage. *Applied Sciences*, 9(4), 655.

[24] Shang, J., Liu, J., Chen, Z., McNairn, H., & Davidson, A. (2022). Recent advancement of synthetic aperture radar (SAR) systems and their applications to crop growth monitoring. IntechOpen.

[25] Kushwaha, A., Dave, R., Kumar, G., Saha, K., & Khan, A. (2022). Assessment of rice crop biophysical parameters using Sentinel-1 C-band SAR data. *Advances in Space Research*, 70(12), 3833–3844.

[26] Liu, Z., et al. (2026). In-season crop yield prediction: State of the art and future research direction. *International Journal of Applied Earth Observation and Geoinformation*, 146, 105129.

[27] Amazirh, A., et al. (2026). Seasonal field-scale wheat yield forecasting using XGBoost with radar, optical, and weather data in Morocco. *International Journal of Applied Earth Observation and Geoinformation*, 149, 105242.
[28] Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1–2), 195–213.

[29] Pettorelli, N., et al. (2005). Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends in Ecology & Evolution*, 20(9), 503–510.
[30] Chen, J. M., & Black, T. A. (1992). Defining leaf area index for non-flat leaves. *Plant, Cell & Environment*, 15(4), 421–429.

[31] Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2), 127–150.

[32] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *ACM SIGKDD*, 785–794.
[33] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *NeurIPS*, 30, 3146–3154.
[34] Guo, F., et al. (2024). A novel transformer-based neural network under model interpretability for improving wheat yield estimation. *Computers and Electronics in Agriculture*, 223, 109111.

[35] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

[36] Song, C., et al. (2025). Wheat yield prediction based on parallel CNN-LSTM-attention with transfer learning model. *Agriculture*, 15(23), 2519.

[37] Wu, X., et al. (2026). TCN–LSTMs: A robust and interpretable framework for winter wheat yield prediction using multi-source time-series data. *International Journal of Applied Earth Observation and Geoinformation*, 149, 105274.

[38] Baatz, R. (2025). Generalization and feature attribution in machine learning models for crop yield and anomaly prediction in Germany. *arXiv:2512.15140*.

[39] Yenkikar, A., Mishra, V. P., Bali, M., & Ara, T. (2025). An explainable AI-based hybrid machine learning model for interpretability and enhanced crop yield prediction. *MethodsX*, 15, 103442.

[40] Copernicus Climate Change Service (C3S). (n.d.). ERA5-Land climate data. Retrieved June 2026. https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview

[41] European Space Agency (ESA). (n.d.). Sentinel-1 satellite mission. Retrieved June 2026. https://sentinel.esa.int/web/sentinel/missions/sentinel-1

[42] European Space Agency (ESA). (n.d.). Sentinel-2 satellite mission. Retrieved June 2026. https://sentinel.esa.int/web/sentinel/missions/sentinel-2

[43] Ministère de l'Agriculture et du Développement Rural (MADR), Algérie. (2020). Série B : Statistiques Agricoles (2016–2019). Direction des Statistiques Agricoles.

[44] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE. *PeerJ Computer Science*, 7, e623.

[45] Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, 1–10.

[46] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79–82.

[47] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin.

[48] De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, 192, 38–48.

[49] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[50] McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, 445(1), 51–56.

[51] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95.

[52] Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.

[53] Abadi, M., Agarwal, A., Barham, P., et al. (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*.

[54] Chollet, F. (2015). Keras: The Python deep learning library. Retrieved June 2026. https://keras.io

[55] Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2), 22–30.

[56] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NeurIPS*, 30, 4765–4774.

[57] Joblib Developers. (2010). Joblib: Running Python functions as pipeline jobs. Retrieved June 2026. https://joblib.readthedocs.io

[58] Streamlit Inc. (2018). Streamlit: The fastest way to build and share data apps. Retrieved June 2026. https://docs.streamlit.io

[59] Carneiro, T., Da Nóbrega, R. V. M., Nepomuceno, T., Bian, G. B., De Albuquerque, V. H. C., & Reboucas Filho, P. P. (2018). Performance analysis of Google Colaboratory as a tool for accelerating deep learning applications. *IEEE Access*, 6, 61677–61685.

[60] Microsoft. (n.d.). Visual Studio Code documentation. Retrieved June 2026. https://code.visualstudio.com/docs