



INSTITUT NATIONAL SUPÉRIEUR D'INFORMATIQUE

Mémoire de Projet en vue de l'obtention du Master 1

Domaine : Sciences et Technologies

Mention : Informatique

Parcours : Intelligence Artificielle et Data Science

Prévision des Prix des Produits de Première Nécessité à Madagascar par Apprentissage Profond

Présenté par :

ANDRIAMAMY Rova Isaia

Sous la direction de :

Dr. RAMASONDRANO Andriamanjaka

Année Universitaire : 2024-2025

Résumé

Cette étude propose une approche d'apprentissage profond pour la prévision des prix de produits agricoles et non agricoles à Madagascar, Elle inclut le riz (local, importé, Malokila), le manioc (frais et sec), le maïs et le maïs concassé, le sucre (blanc et brun), l'huile végétale, le fuel diesel/essence et la farine de blé. Basée sur l'analyse comparative de sept architectures (LSTM, GRU, RNN, Random Forest, XGBoost, Deep DNN, LSTM-RNN), la recherche démontre la supériorité du modèle LSTM pour capturer les dépendances temporelles complexes. Le modèle optimal, entraîné sur des données historiques de 2004 à 2023 couvrant 21 régions et 14 produits, utilise des séquences de 6 mois pour prédire le prix du mois suivant.

Les résultats obtenus après un prétraitement rigoureux montrent des performances exceptionnelles : MAPE de 4,85%, MAE de 126,03 MGA et R^2 de 0,9451. Des variations de précision sont observées selon les produits, avec des performances remarquables pour l'huile végétale ($R^2=0,98$), le sucre ($R^2=0,96$) et le riz local ($R^2=0,95$). Une interface web interactive a été développée pour valider le modèle et faciliter son utilisation opérationnelle.

Ce travail constitue la première initiative exhaustive de prévision systématique des prix à Madagascar et offre un cadre méthodologique reproductible pour les décideurs et acteurs économiques. Les applications potentielles incluent l'optimisation des chaînes d'approvisionnement et la stabilisation des marchés.

Mots-clés : Prévision de prix, LSTM, Apprentissage profond, Série temporelle, Madagascar, Interface web, Produits agricoles et non agricoles.

Table des matières

Liste des Figures et Tableaux	5
Liste des Abréviations et Acronymes	6
1 Introduction	7
1.1 Contexte Général	7
1.2 Problématique et Justification	8
1.3 Objectifs	9
1.4 Structure du Rapport	9
2 Revue de Littérature	9
2.1 Économie des Prix à Madagascar	9
2.2 Approches en Machine Learning et Deep Learning	11
2.3 Hypothèses de Recherche	13
2.4 Lacunes et Positionnement	13
2.4.1 Synthèse des Lacunes Identifiées	14
2.4.2 Positionnement et Contribution du Projet	14
3 Méthodologie	15
3.1 Conception de l'Étude	15
3.2 Collecte et Description des Données	15
3.2.1 Feature Engineering	17
3.2.2 Analyse de Corrélation et Sélection des Variables	18
3.3 Prétraitement des Données	20
3.3.1 Gestion des Valeurs Aberrantes	20
3.3.2 Pipeline de Prétraitement	20
3.4 Évaluation Comparative des Modèles	22
3.4.1 Modèles Évalués	22
3.4.2 Protocole d'Évaluation Comparative	26
3.5 Architecture du Modèle LSTM Optimal	27
3.5.1 Composants Architecturaux	27
3.5.2 Configuration,Évaluation et Entraînement	28
3.5.3 Métriques d'Évaluation	28
3.5.4 Avantages de l'Architecture LSTM	29
3.5.5 Validation et Déploiement	29
3.5.6 Pipeline Méthodologique de LSTM	30

4	Résultats	31
4.1	Analyse Descriptive	31
4.1.1	Structure du jeu de données	31
4.1.2	Transformation de la variable cible	32
4.2	Performance Comparative des Modèles	33
4.2.1	Analyse des Résultats Comparatifs	33
4.3	Performance Globale du Modèle LSTM Optimal	34
4.4	Performances par Produit	35
4.5	Architecture du Système de Prédiction de Prix des Produits	35
4.5.1	Système d'Authentification	36
4.5.2	Import et Analyse de Fichiers	36
4.5.3	Prédiction Simple	37
4.5.4	Prédiction Batch	38
4.5.5	Section d'Aide	39
4.5.6	Prédiction Test du Système	39
4.6	Fonctionnement Technique du Système	40
4.6.1	Architecture LSTM Implémentée	40
4.6.2	Workflow de Prédiction	41
4.6.3	Gestion des Données	41
4.6.4	Performance et Validation	41
5	Discussion	41
5.1	Interprétation des Résultats	41
5.1.1	Validation des Hypothèses de Recherche	41
5.1.2	Analyse des Corrélations Économiques	42
5.1.3	Analyse de la Performance par Produit	42
5.2	Comparaison avec la Littérature	43
5.3	Limites	44
5.4	Implications Pratiques et Éthiques	45
5.4.1	Applications Opérationnelles	45
5.4.2	Considérations Éthiques	45
6	Conclusion et Perspectives	45

Liste des Figures et Tableaux

Figures

Figure 1 : Matrice de corrélation des variables avant traitement de la multicollinéarité.

Figure 2 : Architecture du modèle LSTM optimal pour la prévision des prix

Figure 3 : Architecture du pipeline méthodologique de modèle LSTM

Figure 4 : Matrice de corrélation après élimination des variables fortement corrélées.

Figure 5 : Évolution des courbes de loss (MSE) lors de l'entraînement du modèle LSTM

Figure 6 : Interface de connexion - Page d'accueil de l'application

Figure 7 : Interface principale - Import et Analyse des données

Figure 8 : Interface de prédiction simple - Analyse ciblée

Figure 9 : Interface de prédiction batch - Traitement en masse

Figure 10 : Section d'aide - Documentation complète du système

Figure 11 : Comparaison des prix prédits et réels pour Farine de blé (Wheat Flour)

Tableaux

Tableau 1 : Sources et description des variables explicatives.

Tableau 2 : Distribution des observations par produit dans le jeu de données complet.

Tableau 3 : Impact de la winsorization sur la distribution des prix

Tableau 4 : Comparaison des performances des différents modèles de prédiction

Tableau 5 : Performances par produit

Tableau 6 : Comparaison des prix prédits et réels pour Farine de blé (Wheat Flour) (2023)

Liste des Abréviations et Acronymes

AGOA	Loi sur la croissance et les opportunités en Afrique (African Growth and Opportunity Act)
BAD	Banque Africaine de Développement
CLA	Chenille Légionnaire d'Automne (Fall Armyworm)
DGI	Direction Générale des Impôts
DGD	Direction Générale des Douanes
DNN	Réseau de Neurones Profond (Deep Neural Network)
EPASA	Évaluation de la Production Agricole et de la Sécurité Alimentaire
FAO	Organisation des Nations Unies pour l'alimentation et l'agriculture
FFI	Flux Financiers Illicites
FMI	Fonds Monétaire International
GRU	Unité Récurrente à Porte (Gated Recurrent Unit)
HDX	Humanitarian Data Exchange
IA	Intelligence Artificielle
IDE	Investissement Direct Étranger
INSTAT	Institut National de la Statistique de Madagascar
JIRAMA	Société nationale d'électricité et d'eau de Madagascar
LSTM	Mémoire à Long Court Terme (Long Short-Term Memory)
MAE	Erreur Absolue Moyenne (Mean Absolute Error)
MAEP	Ministère de l'Agriculture et de l'Élevage
MAPE	Erreur en Pourcentage Absolue Moyenne (Mean Absolute Percentage Error)
MECIE	Mise en Compatibilité des Investissements avec l'Environnement
MEF	Ministère de l'Economie et des Finances
MGA	Ariary malgache (Monnaie nationale)
MSE	Erreur Quadratique Moyenne (Mean Squared Error)
ODD	Objectifs de Développement Durable
OMH	Office Malgache des Hydrocarbures
PGE	Politique Générale de l'État
PIB	Produit Intérieur Brut
PPP	Partenariat Public-Privé
R²	Coefficient de Détermination
RMSE	Racine de l'Erreur Quadratique Moyenne (Root Mean Square Error)
RNN	Réseau de Neurones Récurrent (Recurrent Neural Network)
SADC	Communauté de Développement de l'Afrique Australe
SPI	Indice de Précipitations Standardisé (Standardized Precipitation Index)
USAID	Agence des États-Unis pour le développement international
USD	Dollar américain (United States Dollar)
VAR	Modèle Vectoriel Autoregressif (Vector Autoregression)
VECM	Modèle à Correction d'Erreur Vectoriel (Vector Error Correction Model)
WFP	Programme Alimentaire Mondial (World Food Programme)
XGBoost	Extreme Gradient Boosting
ZLECAf	Zone de Libre-Échange Continentale Africaine

1 Introduction

1.1 Contexte Général

Madagascar, un État insulaire d’Afrique australe couvrant 587 041 km² avec une population estimée à plus de 31 millions d’habitants (dont 64,8 % ont moins de 25 ans), reste classé parmi les pays à faible revenu, avec un PIB par habitant de 549 USD en 2024. L’économie malgache est hautement vulnérable aux chocs climatiques et aux fluctuations mondiales, dépendant fortement des prix stables de produits essentiels [13]. Ce rapport se concentre sur la prévision des prix pour une gamme élargie de produits : agricoles comme le riz local (aliment de base), le manioc frais/sec et le maïs ; non agricoles comme le sucre blanc/brun, l’huile végétale, le fuel diesel/essence et la farine de blé. Ces prix, exprimés en MGA/kg ou L, varient selon 21 régions (ex. : Analamanga, Atsimo Andrefana, etc.) et sont influencés par des facteurs comme les précipitations, l’inflation, les importations/exportations et les événements extrêmes (cyclones, feux). L’agriculture, la foresterie et la pêche représentent environ 25 % du PIB et 75 % de l’emploi total, constituant un pilier économique essentiel mais sous-valorisé dans les comptes nationaux [13].

Historiquement, la production agricole illustre cette volatilité : en 2018-2019, la production de riz paddy a atteint 3,91 millions de tonnes (+17,8 % vs. 2017-2018, +9 % vs. moyenne 2014-2018), grâce à une pluviométrie favorable (moyenne 1 513 mm/an, supérieure à la normale dans le Nord/Sud-Ouest) et l’absence de chocs majeurs, sauf dans Diana et Atsimo Atsinanana. Le maïs a augmenté de 7,8 % à 219 220 tonnes, malgré des pertes de 47 % dues aux chenilles légionnaires d’automne (CLA) dans des régions comme Ihorombe (-38 %), Atsimo Andrefana (-29 %), Atsimo Atsinanana (-17 %), Amoron’i Mania (-15 %), Alaotra Mangoro (-8 %). Le manioc a progressé de 12,25 % à 2,91 millions de tonnes, particulièrement dans le Sud (Androy +73 %, Anosy +50 %, Atsimo Andrefana +41 %). Ces dynamiques ont réduit les besoins en importations de riz à 340 000 tonnes pour 2019/2020, mais les prix du riz restent saisonniers, avec des hausses de 20 % entre novembre et mai/juin, exacerbées par une capacité de stockage limitée et des pertes post-récolte de 15-20 % [14].

En 2024, la croissance économique globale est estimée à 4,2 % (stable par rapport à 2023), tirée par les performances agricoles, le textile, le tourisme et les télécommunications, malgré une contraction minière de 20,8 % due à la baisse de production de nickel et cobalt. Les perspectives pour 2025-2026 prévoient une croissance de 3,8 % puis 4,0 %, soutenue par l’agriculture et les industries extractives, mais menacée par les chocs climatiques (cyclones, sécheresses dans le Sud) et les tensions géopolitiques (hausse des tarifs douaniers US, représentant 47 % sur les exportations malgaches vers les États-Unis). L’inflation a reculé à 7,6 % en 2024 (contre 9,9 % en 2023), principalement due à la hausse des coûts de production, avec des projections à 7,0 % en 2025. La pauvreté reste alarmante : 75 % au niveau national

en 2022 (79,9 % rural, 55,5 % urbain, >91 % dans le Sud à cause notamment des chocs climatiques), exacerbée par ces volatilités, avec plus de 80 % de la population sous le seuil de 2,15 USD/jour [13]. En 2018-2019, l'insécurité alimentaire a touché 200 000 personnes en phase sévère et 1,9 million en modérée, avec des améliorations dans le Sud grâce à de bonnes récoltes, mais des poches persistantes à Bekily, Amboasary, Ambovombe et Ampanihy. Les stocks des ménages s'épuisent rapidement (majorité <2 mois post-récolte), menant à une confiance faible (70 % incertains/pessimistes pour sept-déc. 2019), amplifiée par des contraintes comme l'irrigation limitée (1,2 M ha en maîtrise totale) et faible utilisation d'intrants (5 kg/ha engrais) [14].

Les exportations ont chuté de 16,8 % en 2024 (vanille, girofle, nickel), tandis que les importations (énergie, biens de consommation comme le riz) ont augmenté de 3,4 %, creusant le déficit courant à 5,0 % du PIB. La baisse des exportations, conjuguée à une progression des importations, a aggravé le déficit du compte courant à 5,0 % du PIB. Alors la prévision précise est cruciale pour les acteurs locaux : agriculteurs pour l'optimisation des récoltes, importateurs pour la gestion des stocks pour éviter la dépendance des importations pour chaque année, et autorités pour les subventions ciblées, alignées sur l'objectif d'autosuffisance rizicole d'ici 2030 et pour les autres produits [13]. Ce projet pionnier en prévision de prix de produits à Madagascar vise à combler un vide analytique en intégrant ces dynamiques historiques et régionales.

1.2 Problématique et Justification

La volatilité des prix des produits à Madagascar constitue un défi majeur pour la planification économique et la sécurité alimentaire. Cette instabilité persistante est amplifiée par plusieurs facteurs interdépendants : la vulnérabilité accrue aux chocs climatiques, la fluctuation des importations et exportations, les coûts de transport élevés au niveau national, les ravageurs et maladies agricoles, ainsi que la baisse des cours mondiaux des matières premières, comme documenté dans l'étude EPASA 2019 [14].

La problématique centrale réside dans l'absence d'outils prédictifs sophistiqués permettant d'anticiper l'évolution des prix des produits essentiels. Les approches existantes se limitent principalement à des analyses statistiques descriptives pour constater *a posteriori* les hausses de prix et évaluer l'insécurité alimentaire, sans capacité de prévision proactive permettant une intervention anticipée [9]. Ce déficit analytique se traduit par une gestion réactive plutôt que proactive des crises alimentaires, particulièrement durant les périodes de soudure où les hausses de prix atteignent fréquemment 20% [14].

Face à ce constat, ce travail propose un modèle de prédiction basé sur les réseaux de neurones LSTM (*Long Short-Term Memory*), spécialement conçu pour capturer les dynamiques temporelles complexes des marchés malgaches. Le modèle, spécifiquement calibré pour le contexte local, traite des séquences temporelles de 6 mois et intègre des métadonnées régionales et par produit. Il incorpore l'ensemble des variables déterminantes identifiées dans la

littérature - incluant les facteurs climatiques, économiques, logistiques et saisonniers - offrant ainsi une approche prédictive innovante pour anticiper les tendances des marchés et soutenir la prise de décision des acteurs économiques et institutionnels à Madagascar.

La justification de cette approche repose sur la nécessité urgente de transformer la compréhension descriptive des déterminants des prix, bien documentée dans les rapports existants, en une capacité opérationnelle de prévision quantitative. Cette innovation méthodologique vise directement à renforcer la résilience du système alimentaire malgache face aux chocs récurrents et à soutenir l'objectif national d'autosuffisance rizicole d'ici 2030.

1.3 Objectifs

Développer un modèle de prévision des prix des produits robuste et fiable, adapté aux données multivariées malgaches pour les décideurs agricoles pour améliorer la sécurité alimentaire d'ici dans l'année prochaine.

1.4 Structure du Rapport

Ce rapport est organisé en six chapitres principaux. L'**Introduction** présente le contexte et la problématique. La **Revue de Littérature** examine les travaux antérieurs et positionne cette recherche. La **Méthodologie** détaille la collecte, prétraitement des données et l'architecture du modèle. Les **Résultats** présentent les performances obtenues. La **Discussion** interprète ces résultats et les confronte à la littérature. Enfin, la **Conclusion** synthétise les apports et propose des perspectives.

2 Revue de Littérature

2.1 Économie des Prix à Madagascar

Le **Rapport Pays 2025 de la Banque Africaine de Développement (BAD)** fournit un cadre analytique essentiel pour comprendre l'environnement macroéconomique dans lequel s'inscrit la formation des prix agricoles à Madagascar. Sur le plan méthodologique, ce rapport synthétise des données secondaires officielles provenant des institutions nationales (**INSTAT, MEF**), du Fonds Monétaire International (**FMI**) et de la Banque Mondiale, offrant une vue d'ensemble des tendances pour la période **2020-2026**.

Le rapport met en lumière plusieurs facteurs clés influençant les prix. Il souligne la vulnérabilité de l'économie aux chocs externes, telle que la hausse projetée de **15%** du prix du diesel en **2025** – influencée par les tensions géopolitiques – qui impacte directement les coûts de production et de transport des produits agricoles et non-agricoles. Cette dynamique s'inscrit dans un contexte inflationniste persistant, avec un taux d'inflation de **7,6%** en **2024**, que la Banque Centrale tente de contenir par une politique monétaire restrictive. Parallèlement, le rapport confirme l'importance des facteurs saisonniers, avec des hausses pouvant atteindre **20%** pour le riz en saison sèche.

De plus, il identifie la forte dépendance aux exportations de produits de base comme un risque majeur, illustré par la baisse de **16,8%** des exportations en **2024** pour des produits à haute valeur comme la vanille (représentant **8%** du PIB) et le girofle. Cette vulnérabilité contraste avec la contribution vitale de l'agriculture à l'économie nationale (**25%** du PIB, **75%** de l'emploi) et la croissance économique soutenue de **4,2%** enregistrée en **2024**.

Le contexte **socio-économique**, caractérisé par une pauvreté rurale affectant **79,9%** de la population, des inégalités importantes (indice de Gini de **36,8**) et un taux de malnutrition de **39,7%** – soit le double de la moyenne africaine en **2022** – est identifié comme un amplificateur critique de l'impact des chocs de prix sur la sécurité alimentaire.

Les principaux **points forts** de cette étude sont son analyse holistique reliant les dynamiques macroéconomiques, sectorielles et sociales, et sa perspective prospective. Ses **limita-tions** résident dans son niveau d'analyse agrégé (national), qui peut occulter les dynamiques locales des marchés, et son caractère principalement qualitatif, qui ne quantifie pas l'élasticité des prix par rapport à ses différents déterminants [13].

L'Évaluation de la Production Agricole et de la Sécurité Alimentaire (EPASA) de 2019 constitue une étude de référence qui illustre de manière significative les dynamiques historiques et structurelles du secteur agricole malgache. Sur le plan méthodologique, ce rapport s'appuie sur une **enquête terrain exhaustive** couvrant 21 des 22 régions, combinant des **questionnaires standardisés**, des **entretiens avec des acteurs clés** (agriculteurs, commerçants, autorités) et une **analyse de données secondaires** (statistiques officielles, images satellites, données de l'Observatoire du Riz). Son principal point fort réside dans son **approche holistique** et sa rigueur, utilisant des outils reconnus comme l'indice de précipitations standards (SPI) et le cadre d'analyse CARI pour la sécurité alimentaire, ce qui lui confère une grande valeur descriptive.

Les données de la campagne 2018-2019 sont éloquentes : la production rizicole (paddy) a connu une augmentation notable de **17,8 %**, atteignant **3,91 millions de tonnes**, principalement attribuable à des conditions pluviométriques favorables. En revanche, la filière maïs a subi des pertes considérables, avec une baisse de rendement de **47 %** due aux dégâts de la Chenille Légionnaire d'Automne (CLA) qui a touché sévèrement cinq régions productrices, notamment la région Ihorombe (**-38 %**), Atsimo Andrefana (**-29%**), Atsimo Atsinanana (**-17%**), Amoron'i Mania (**-15%**) et Alaotra Mangoro (**-8%**). Paradoxalement, la production de manioc a enregistré une progression remarquable de **12,25 %** dans le Sud, performance qui a contribué à réduire les besoins d'importation de riz à **340 000 tonnes** pour la campagne 2019/2020.

Cependant, le rapport présente également des **limites importantes** qui justifient le recours à des méthodes prédictives comme l'IA. Son approche est essentiellement *descriptive et a posteriori*, avec une fréquence annuelle qui ne capture pas les dynamiques intra-annuelles des prix. De plus, il souffre de l'**absence de modélisation multivariée** pour quantifier les interactions complexes entre les facteurs, et repose partiellement sur des **données déclaratives** susceptibles de biais.

Malgré des améliorations ponctuelles de la production, la situation alimentaire demeurait ainsi préoccupante avec **200 000 personnes** en insécurité alimentaire sévère et **1,9 million** en insécurité modérée. Ces chiffres mettent en évidence plusieurs caractéristiques structurelles : une saisonnalité marquée des prix du riz (pic en novembre, creux en juin), des contraintes techniques persistantes (faible utilisation d'intrants avec seulement **5 kg d'engrais par hectare**), et une volatilité des prix atteignant **20 %**. Les importations céréalières restent significatives avec une moyenne de **380 000 tonnes par an** (2014-2019), et le bilan 2019/2020 projetait un déficit céréalier total de **531 000 tonnes**.

Le rapport EPASA 2019 fournit un **diagnostic contextuel inestimable** pour comprendre les déterminants des prix, ses limitations méthodologiques soulignent la nécessité de le compléter par des approches capables d'intégrer la multiplicité des facteurs en jeu et de produire des prévisions opérationnelles. [14].

2.2 Approches en Machine Learning et Deep Learning

Des travaux récents sur les marchés émergents, **Chen et al. (2021) - Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques en malaisie** représente une avancée significative dans le domaine de la prédiction des prix agricoles par l'intelligence artificielle. Les chercheurs ont réalisé une comparaison systématique de cinq algorithmes de pointe - **ARIMA, SVR, Prophet, XGBoost et LSTM** - sur des données historiques couvrant la période **2008-2020** et incluant des produits de base essentiels comme le **poulet**, le **piment** et la **tomate**. Leur approche méthodologique a comporté deux phases expérimentales distinctes : une analyse **univariée** basée exclusivement sur les séries temporelles de prix, suivie d'une approche **multivariée** enrichie par l'ajout de variables externes critiques incluant la **température**, l'**humidité**, les **précipitations** et le **prix du pétrole**.

Les résultats obtenus révèlent des performances différentielles selon la complexité des données. Dans la configuration **univariée**, le modèle **ARIMA** démontre une performance optimale avec un **MSE** moyen de **0,251**, surpassant initialement le modèle **LSTM**. Cependant, dans le cadre **multivarié** enrichi, le **LSTM** domine nettement l'ensemble des algorithmes avec un **MSE** moyen de **0,304**, représentant une amélioration remarquable de **45,5%** par rapport à sa performance initiale. Cette supériorité s'explique par la capacité intrinsèque du **LSTM** à modéliser les relations non-linéaires complexes et les dépendances temporelles à long terme, tout en contournant efficacement le problème de disparition du gradient.

Les **principaux atouts** de cette recherche résident dans son design expérimental complet intégrant une validation progressive, l'implémentation d'**un système web automatisé utilisant Django et TensorFlow**, et l'exploitation de données réelles couvrant **12 années** d'observations continues. Néanmoins, l'étude présente certaines **limitations** dont un volume de données restreint pour les modèles de deep learning, un nombre limité de **variables ex-**

ternes intégrées dans l'analyse **multivariée**, et une applicabilité principalement circonscrite au contexte spécifique **malaisien** sans validation transnationale. [15].

Mohamed et al. (2022) abordent spécifiquement la prévision de prix d'**articles saisonniers**. Cette recherche égyptienne explore un créneau spécialisé mais crucial : la prédiction des prix des **articles saisonniers**. L'étude s'appuie sur un jeu de données substantiel de **20 179 enregistrements** d'entraînement provenant du commerce électronique, procédant à une comparaison méthodique entre **SVR**, la **régression linéaire**, **Random Forest**, **Ridge** et **ARIMA**.

Les résultats démontrent de manière éloquente la supériorité de l'algorithme **Random Forest**, atteignant des performances exceptionnelles avec un MAPE de **1,63%**, un **R²** de **61%** et un RMSE de **31,31**. Le modèle **ARIMA** se positionne comme deuxième meilleur performeur avec un MAPE de **7,8%** et un RMSE de **44,35**, confirmant ainsi la pertinence persistante des approches statistiques traditionnelles dans des contextes saisonniers. Ces résultats valident l'efficacité supérieure des méthodes par ensembles pour capturer les patterns saisonniers complexes.

Les **forces majeures** de cette recherche incluent son innovation thématique en ciblant spécifiquement les **produits saisonniers**, son évaluation multicritère exhaustive couvrant quatre métriques distinctes, et son implémentation reproductible utilisant **Python/Scikit-learn**. Cependant, l'étude présente certaines **limitations** dont un ensemble de variables d'entrée restreint comprenant seulement **10 caractéristiques partiellement anonymisées**, l'utilisation majoritaire de paramètres par défaut pour les modèles, et un cadre d'application limité aux **cadeaux de Noël** qui pourrait restreindre la généralisation des résultats. [16].

Brignoli et al. (2024) ont mené une étude comparative approfondie sur la prédiction des prix à terme du **maïs** en utilisant des **LSTM-RNN** (Long Short-Term Memory Recurrent Neural Networks). Leur recherche s'appuie sur des données quotidiennes de **2000 à 2020** et intègre un large éventail de variables explicatives incluant des prix de **produits agricoles (blé dur, blé tendre, soja, lait)**, des produits énergétiques (**pétrole brut WTI, essence**) et des indices financiers.

La **méthodologie** employée compare systématiquement les performances des **LSTM** avec des modèles économétriques traditionnels (**VAR, VECM et ARIMA**), en testant différentes techniques de prétraitement des données : **Full Adjustment (FA)** combinant transformation logarithmique, **désaisonnalisation** et **détendance via LOESS**, et **First Difference (FD)** pour la stationnarisation.

Les **résultats** démontrent la supériorité des **LSTM-RNN**, particulièrement sur les horizons de prévision longs (**30 jours**) où ils affichent un MAE inférieur de **10%** aux meilleurs modèles classiques. Cette performance est encore plus marquée lorsque les modèles sont entraînés sur l'ensemble des données disponibles, les **LSTM** montrant une capacité remarquable à gérer automatiquement les ruptures structurelles sans nécessiter de détection manuelle préalable - un avantage décisif par rapport aux modèles **VAR/VECM** qui requièrent une identification explicite des points de rupture.

Sur les horizons courts (**7 jours**), les **LSTM** maintiennent leur compétitivité avec des **MAE** autour de **2,5** contre **5+** pour les **VAR** sur données **FA**, bien que leur performance dépende crucialement du prétraitement appliqué. L'étude révèle que l'ajustement **FD** s'avère plus efficace que l'approche **FA** pour les séries de **prix agricoles**, réduisant significativement les erreurs de prédiction.

L'architecture optimisée des **LSTM** (**5 couches, 333 nœuds, régularisation =0,1**) confirme leur robustesse pour capturer les dynamiques non-linéaires complexes des marchés agricoles. Ces conclusions s'alignent parfaitement avec notre problématique malgache, où la volatilité accrue (**chocs climatiques, disruptions géopolitiques**) nécessite des modèles capables de s'auto-adapter aux changements structurels - une capacité intrinsèque des **LSTM** démontrée empiriquement par Brignoli et al.

La principale **limite** identifiée réside dans le caractère "boîte noire" des **LSTM** et leur dépendance aux hyperparamètres, bien que l'optimisation bayésienne ait permis de mitiger ce dernier inconvénient. Néanmoins, l'étude établit solidement la pertinence des approches **LSTM** pour la prévision des prix agricoles dans des contextes volatils.[17].

2.3 Hypothèses de Recherche

Cette étude repose sur les hypothèses scientifiques suivantes, formulées à partir de la revue de littérature et des spécificités du contexte malgache :

- **H1** : Les variables temporelles (précipitations, inflation, saisonnalité, prix petrol) expliquent significativement la variance des prix des produits, conformément aux résultats de Chen et al. (2021), la BAD (2025) et Brignoli et al. (2024) sur l'importance des facteurs exogènes dans les marchés émergents
- **H2** : L'architecture LSTM surpasse significativement les modèles traditionnels pour capturer les séquences temporelles complexes, comme démontré dans les travaux de Brignoli et al. (2023) et Chen et al. (2021) sur les futures agricoles
- **H3** : L'intégration des métadonnées régionales et des caractéristiques produits améliore la précision des prévisions, en cohérence avec les constats de l'EPASA (2019) sur l'hétérogénéité territoriale malgache
- **H4** : Le modèle permet d'identifier les facteurs clés influençant la volatilité des prix à Madagascar, répondant au besoin d'outils analytiques identifié par la BAD (2025) pour la stabilisation des marchés

2.4 Lacunes et Positionnement

Nos revues de littérature mettent en lumière un paysage de recherche caractérisé par des diagnostics approfondis mais un déficit prononcé en outils prédictifs opérationnels pour le contexte malgache. Cette recherche se positionne pour combler plusieurs lacunes identifiées dans la littérature existante, notamment le peu d'études intégrant simultanément les dimensions régionales et produits spécifiques au contexte malgache pour la prévision de prix de

produits.

2.4.1 Synthèse des Lacunes Identifiées

L’analyse des études de référence révèle trois limitations principales :

- **Déficit Prédictif des Études Existantes** : Les rapports du BAD (2025) et de l’EPASA (2019) offrent un diagnostic contextuel et rétrospectif de haute qualité. L’étude EPASA 2019 met particulièrement en évidence un déficit important : l’absence de modèles prédictifs capables d’anticiper les hausses de prix pendant la période de soudure (ex : septembre-décembre), et ce malgré la réalisation d’enquêtes terrain approfondies couvrant 21 régions.
- **Inadéquation des Modèles au Contexte Malgache** : Les travaux de Chen et al. (2021) et Brignoli et al. (2024) démontrent la supériorité des modèles de *Deep Learning* (LSTM) pour la prédiction de prix agricoles. Cependant, ces modèles sont calibrés pour des économies émergentes structurées (Malaisie) ou des marchés à terme internationaux, sans prise en compte des spécificités de l’économie malgache.
- **Manque d’Intégration des Déterminants Locaux** : Aucune des études examinées n’intègre de manière exhaustive l’ensemble des variables critiques identifiées pour Madagascar, notamment les vulnérabilités structurelles documentées telles que les stocks limités à moins de deux mois et les situations d’insécurité alimentaire modérée [14].

2.4.2 Positionnement et Contribution du Projet

Ce projet pionnier en prévision des prix vise précisément à combler ces lacunes, en s’appuyant sur une analyse comparative des études existantes sur Madagascar et des recherches internationales similaires qui confirme la pertinence des variables explicatives sélectionnées dans notre jeu de données d’entraînement. Sa contribution est triple :

1. **Avancée Méthodologique** : Il propose le premier modèle de prédiction des prix agricoles à Madagascar basé sur l’*Apprentissage profond*, spécifiquement conçu et entraîné sur des données locales. Cette validation externe renforce la légitimité de notre approche et justifie le développement d’un modèle d’intelligence artificielle spécifiquement conçu pour appuyer la prise de décision des acteurs économiques et institutionnels à Madagascar.
2. **Innovation Contextuelle** : Le modèle est conçu pour capturer les vulnérabilités structurelles de l’économie malgache en intégrant explicitement dans son analyse les variables de **vulnérabilité socio-économique** comme amplificateurs de chocs, une dimension absente des modèles génériques de la littérature.
3. **Opérationnalité et Granularité** : En s’appuyant sur des données infra-annuelles et en produisant des prévisions à l’échelle régionale, le projet surmonte les limites

d'agrégation nationale des rapports existants. Il fournit ainsi des *insights* actionnables pour les acteurs locaux, des pouvoirs publics aux coopératives agricoles.

En synthèse, ce projet comble une lacune critique en transformant la compréhension descriptive des déterminants des prix, établie par la littérature existante, en une capacité prédictive quantitative et contextualisée. Il représente une évolution nécessaire des méthodologies, passant de l'analyse *a posteriori* à la prévision opérationnelle, afin de renforcer la résilience du système alimentaire malgache face à une volatilité accrue.

3 Méthodologie

3.1 Conception de l'Étude

Cette recherche adopte une approche quantitative basée sur une méthodologie expérimentale comparative pour le développement et l'évaluation de modèles de prévision des prix. Le cadre méthodologique suit un processus structuré comprenant la collecte des données, le prétraitement, la modélisation comparative et l'évaluation. Une analyse comparative exhaustive a été menée entre sept architectures de modèles distinctes : LSTM, GRU, Random Forest, XGBoost, Simple RNN, Deep DNN et un modèle hybride LSTM-RNN, afin d'identifier l'approche la plus performante pour le contexte spécifique des données malgaches.

L'étude utilise deux datasets historiques de prix de produits à Madagascar couvrant différentes périodes :

- **Dataset 2004-2022** : Humanitarian Data Exchange (HDX) - 14,540 échantillons [1]
- **Dataset 2023** : Institut National de la Statistique de Madagascar (INSTAT) - 27,482 échantillons [2]

Le jeu de données consolidé comprend initialement 42 022 échantillons, couvrant 21 régions et 24 produits agricoles et non agricoles, incluant la variable cible correspondant à l'historique des prix de chaque produit. Plusieurs étapes de prétraitement ont été appliquées, la vérification de la disponibilité des variables explicatives afin d'éviter les valeurs manquantes, de filtrage basé sur les percentiles 5-95%, ainsi que le traitement des valeurs aberrantes relatives aux prix, qui ont été supprimées du jeu de données.

3.2 Collecte et Description des Données

La collecte des données a impliqué l'agrégation de multiples sources pour construire un ensemble de variables explicatives complètes. Un processus d'acquisition automatisé a été mis en œuvre via des scripts Python dédiés, permettant d'extraire et de structurer les données dans des fichiers CSV normalisés. Le tableau 1 présente l'ensemble des variables collectées :

TABLE 1 – Sources et description des variables explicatives

Variable	Source	Description
Feu, cyclone, nbJour_Cyclone, nbJour_Feu	Data Commons Madagascar - Disasters Dashboard [3]	Valeur moyenne mensuelle des événements climatiques
Essence(MGA), Diesel(MGA), ImportFuel(l/mois), Consom-Fuel(L/mois)	Office Malgache des Hydrocarbures (OMH) [4]	Prix moyens mensuels et volumes d'importation
TempMoyenne, Precipitations	Meteostat [5]	Données climatiques moyennes mensuelles
Heures_Ensoleillement(j/h/mois), Jours_de_Pluie(j/mois)	DonnéesMondiales.com [6]	Indicateurs climatiques mensuels
Terre_Agricole(km ²)	Banque Mondiale [7]	Surface agricole par année
Inflation(%)	Trading Economics [8]	Taux d'inflation mensuel général
Consom(kg/pers/an), ImportP-Prod(Tonnes/an), ExportP-Prod(Tonnes/an)	Ministère de l'Agriculture et de l'Élevage [9]	Bilans alimentaires annuels
Production(Tonnes/an)	FAOSTAT [10]	Production agricole annuelle
NbPopulation	City Population [11]	Données démographiques annuelles
Importation_EnMillion(MGA), Exportation_EnMillion(MGA)	World Integrated Trade Solution [12]	Commerce international annuel

La variable **Coût_Transports(MGA/km)** a été calculée manuellement à l'aide d'une formule spécifique adaptée au contexte malgache, intégrant les principaux facteurs influençant les coûts logistiques à Madagascar :

$$CT = CB_{25T} \times FR \times FV \times PC \quad (1)$$

Où :

- CT : Coût de transport moyen (MGA/km)
- CB_{25T} : Consommation de base d'un camion de transport 25 tonnes (0.35 litre/km) [18]
- FR : Facteur de qualité de route (1.4) - reflétant l'état généralement médiocre des routes malgaches
- FV : Facteur de vitesse de conduite (1.0) - vitesse normale adaptée aux conditions locales
- PC : Prix moyen du carburant (MGA/litre) pour le mois correspondant

Cette modélisation permet de capturer la sensibilité du coût de transport aux fluctuations du prix des carburants, tout en tenant compte des spécificités infrastructurelles du réseau routier malgache.

3.2.1 Feature Engineering

Afin d'améliorer la capacité prédictive du modèle et de capturer des relations économiques et agronomiques sous-jacentes, plusieurs variables dérivées ont été générées à partir des données existantes. Ce processus d'ingénierie des variables permet de transformer les données brutes en caractéristiques plus informatives, facilitant ainsi l'apprentissage des patterns complexes par le modèle. Les nouvelles variables créées sont les suivantes :

— **Productivité :**

$$Productivite_tonnes_par_km2 = \frac{Production}{Terre_Agricole}$$

Cette métrique mesure l'efficacité productive par unité de surface cultivée, exprimant le rendement agricole en tonnes par kilomètre carré. Elle permet de comparer la performance productive entre régions indépendamment de leur superficie agricole totale, identifiant ainsi les zones à fort potentiel de rendement.

— **Demande totale :**

$$Demande_totale_tonnes = \frac{NbPopulation \times Consom}{1000}$$

Cette variable estime la demande agrégée annuelle en convertissant la consommation individuelle (kg/personne/an) en volume total requis pour l'ensemble de la population régionale (tonnes/an). La division par 1000 permet la conversion des kilogrammes en tonnes, offrant une vision macroéconomique des besoins alimentaires régionaux.

— **Ratio offre/demande :**

$$Ratio_Offre_Demande = \frac{Production}{Demande_totale_tonnes}$$

Cet indicateur crucial quantifie l'équilibre structurel entre la production locale et la demande théorique. Un ratio supérieur à 1 indique un surplus productif, tandis qu'un ratio inférieur à 1 révèle une dépendance aux importations. Les divisions par zéro ont été traitées en remplaçant les valeurs infinies par des valeurs manquantes pour préserver l'intégrité statistique.

— **Production par habitant :**

$$Production_par_habitant = \frac{Production}{NbPopulation}$$

Cette variable normalise la production totale par rapport à la taille de la population, exprimée en tonnes par personne. Elle facilite les comparaisons inter-régionales en éliminant l'effet de taille démographique, permettant d'identifier les régions à forte autonomie alimentaire par habitant.

— **Jours sans pluie :**

$$Jours_Sans_Pluie = 30 - Jours_de_Pluie$$

Ce complément météorologique offre une vision symétrique des conditions d'ensoleillement mensuelles. En soustrayant les jours de pluie du total mensuel standard de 30 jours, cette variable capture l'exposition aux conditions sèches, particulièrement importante pour l'analyse des cultures sensibles à la sécheresse et la planification des calendriers agricoles.

— **Saison :**

$$Saison = \begin{cases} 1 & \text{si } Mois \in [11, 12, 1, 2, 3, 4] \quad (\text{saison pluvieuse}) \\ 0 & \text{si } Mois \in [5, 6, 7, 8, 9, 10] \quad (\text{saison sèche}) \end{cases}$$

Cette variable binaire encode la dichotomie saisonnière fondamentale du climat malgache, cruciale pour capturer les cycles agricoles et les variations de prix saisonnières. La saison pluvieuse (novembre à avril) correspond à la période de croissance des cultures principales, tandis que la saison sèche (mai à octobre) influence les stratégies de stockage et les prix de marché.

Ces transformations permettent non seulement d'enrichir l'ensemble de caractéristiques disponibles pour l'apprentissage, mais également d'introduire une sémantique économique et agronomique pertinente pour la prévision des prix dans le contexte spécifique de Madagascar. Chaque variable dérivée a été conçue pour refléter des mécanismes économiques réels tout en maintenant une cohérence statistique avec les données source.

3.2.2 Analyse de Corrélation et Sélection des Variables

Afin d'évaluer les relations linéaires entre les différentes variables explicatives et d'identifier d'éventuels problèmes de multicolinéarité, une analyse de corrélation exhaustive a été réalisée. La multicolinéarité, phénomène où les variables indépendantes présentent des corrélations élevées entre elles, peut en effet compromettre la stabilité des estimateurs et réduire l'interprétabilité du modèle.

La Figure 1 présente la matrice de corrélation complète de l'ensemble des variables avant tout traitement. Cette visualisation permet d'identifier les relations interdépendantes entre les différentes dimensions du dataset, incluant les variables originales et les features dérivées.

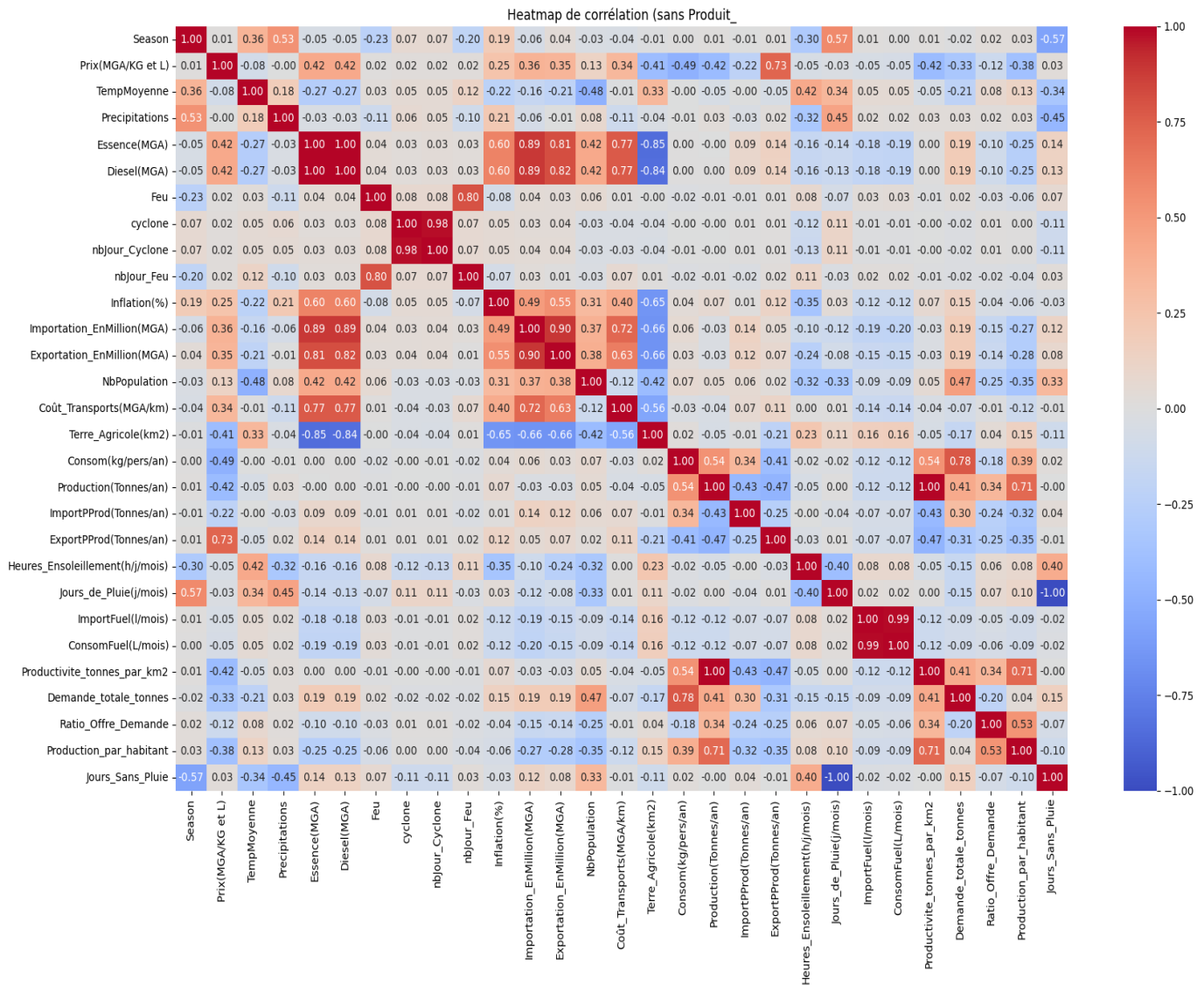


FIGURE 1 – Matrice de corrélation des variables avant traitement de la multicollinéarité

L'analyse révèle plusieurs patterns de corrélation significatifs, notamment des relations fonctionnelles attendues telles que la corrélation positive parfaite entre les prix des carburants ($\rho = +1.0$) et la corrélation négative parfaite entre les indicateurs climatiques complémentaires ($\rho = -1.0$). Ces relations, bien que conceptuellement cohérentes, introduisent une redondance informationnelle préjudiciable à la performance du modèle.

Conformément aux bonnes pratiques en apprentissage statistique, un seuil de corrélation de $|\rho| > 0.85$ a été établi pour identifier et éliminer les variables présentant une multicollinéarité excessive. Cette approche conservatrice permet de préserver la stabilité numérique du modèle tout en maintenant l'essentiel de l'information prédictive.

3.3 Prétraitement des Données

3.3.1 Gestion des Valeurs Aberrantes

L'analyse des valeurs extrêmes a révélé des patterns cohérents avec la réalité économique malgache, nécessitant une approche différenciée selon la nature des variables et leur impact potentiel sur la stabilité du modèle. Pour les variables explicatives, une analyse minutieuse a confirmé que les valeurs extrêmes observées reflétaient des réalités structurelles de l'économie malgache plutôt que des erreurs de mesure.

Concernant la variable cible 'Prix(MGA/KG et L)', une stratégie de traitement différenciée a été mise en œuvre afin de préserver la représentativité des données tout en atténuant l'influence disproportionnée des valeurs extrêmes sur l'apprentissage du modèle. Cette approche duale a été calibrée en fonction de la distribution spécifique de chaque catégorie de produits :

- **Suppression directe des observations aberrantes :** Cette mesure radicale a été réservée aux produits présentant une proportion exceptionnellement élevée de valeurs aberrantes, compromettant la fiabilité des estimations. C'est notamment le cas des huiles végétales conditionnées (*Oil sealed bottle* : 2508 observations aberrantes), des huiles végétales importées (*Oil imported* : 386 observations) et des huiles en vrac (*Oil bulk* : 223 observations), où la prévalence des outliers suggérait des problèmes structurels de collecte ou de marché.
- **Application d'une winsorization à 1% :** Pour les autres produits agricoles de base présentant un nombre modéré de valeurs extrêmes, une approche conservatrice de winsorization a été préférée. Cette technique permet de limiter l'influence des outliers tout en préservant l'information contenue dans les queues de distribution. Les produits concernés incluent le riz local (*Rice local* : 8 observations), le sucre blanc (*White sugar* : 6 observations), le sucre brun (*Brown sugar* : 5 observations), le maïs broyé (*Maize crushed* : 1 observation), la farine de blé (*Wheat flour* : 1 observation) et le riz importé (*Rice imported* : 1 observation). Le mécanisme de winsorization remplace les valeurs extrêmes par les percentiles correspondants au 1er et 99e centile, préservant ainsi l'intégrité statistique de l'échantillon.

Cette stratégie de traitement différencié a produit des résultats statistiquement significatifs sur la distribution de la variable cible. Cette normalisation de la distribution garantit une meilleure stabilité algorithmique tout en maintenant la représentativité économique des données, condition essentielle pour la robustesse des prévisions générées par le modèle.

3.3.2 Pipeline de Prétraitement

Le pipeline de prétraitement a été conçu selon une architecture séquentielle rigoureuse, visant à transformer les données brutes en un format optimal pour l'apprentissage profond tout en préservant l'intégrité sémantique et temporelle des informations. Chaque étape a

été méticuleusement calibrée pour répondre aux spécificités des données malgaches et aux exigences techniques des nos modèles.

1. **Encodage one-hot des variables catégorielles** : Une transformation binaire a été appliquée aux variables nominales 'produits' et 'régions' afin de les rendre compatibles avec les algorithmes d'apprentissage profond. Cette technique crée des vecteurs binaires de dimension égale au nombre de catégories, où chaque position représente l'appartenance à une catégorie spécifique. L'encodage one-hot élimine toute relation d'ordre artificielle entre les catégories tout en préservant leur identité distinctive, essentielle pour la modélisation des effets fixes régionaux et productifs.
2. **Filtrage statistique des séquences** : Un mécanisme de filtrage basé sur les percentiles 5-95% a été implémenté pour éliminer les séquences dont la valeur cible présente des caractéristiques extrêmes. Cette procédure conservatrice permet d'exclure les observations atypiques tout en préservant la variabilité naturelle des prix. Le choix des percentiles 5 et 95 représente un compromis optimal entre la robustesse statistique et la représentativité économique, évitant à la fois le surapprentissage aux outliers et la perte d'information sur la volatilité réelle des marchés.
3. **Construction des séquences temporelles** : Des séquences historiques de 6 mois consécutifs ont été générées pour chaque combinaison unique région-produit, suivant une approche de fenêtrage glissant. Cette méthodologie permet de capturer les dynamiques temporelles à moyen terme tout en maintenant la cohérence spatiale et sectorielle. Chaque séquence sert de contexte historique pour prédire le prix du mois suivant, reproduisant ainsi le processus décisionnel réel des acteurs économiques qui basent leurs anticipations sur l'évolution récente des marchés.
4. **Transformation logarithmique de la variable cible** : La transformation $y_{log} = \log(y + 1)$ a été appliquée à la variable 'Prix(MGA/KG et L)' pour stabiliser la variance et normaliser la distribution. Cette transformation non linéaire atténue l'effet des variations proportionnelles, rendant la série plus stationnaire et améliorant la convergence lors de l'entraînement. L'addition de l'unité (+1) garantit la définition mathématique pour les valeurs nulles théoriques, préservant ainsi l'intégrité de l'ensemble de données.
5. **Normalisation standardisée des features** : Une standardisation via l'algorithme StandardScaler a été appliquée indépendamment aux deux types de variables :
 - **Features temporelles** : Normalisation séparée pour chaque pas de temps, préservant ainsi l'évolution relative des variables dans le temps tout en assurant une échelle commune entre les différentes métriques.
 - **Features statiques** : Normalisation globale à l'ensemble du dataset afin de garantir une comparaison cohérente entre les régions et les produits, indépendamment de leurs caractéristiques structurelles. Cette procédure est particulièrement pertinente pour des variables telles que la superficie des terres agricoles à Madagascar,

et a été utilisée dans les algorithmes Random Forest, XGBoost, Deep DNN et LSTM-RNN.

Cette approche duale maintient la cohérence temporelle tout en standardisant l'échelle des variables, condition essentielle pour la stabilité numérique des algorithmes de descente de gradient.

6. **Séparation stratifiée des données :** Les jeux de données ont été divisés en ensembles d'entraînement, de validation et de test avec la graine aléatoire `random_state=42` pour assurer la reproductibilité des expériences. La stratification temporelle garantit que les ensembles d'entraînement et de test contiennent des représentations proportionnelles de toutes les périodes historiques, évitant ainsi les biais saisonniers et préservant la structure autocorrélative des séries temporelles.
7. **Sauvegarde systématique des artefacts :** L'ensemble des préprocesseurs, incluant les scalers, les encodeurs, les informations de features et les métadonnées, a été sauvegardé dans un format persistant. Cette pratique garantit non seulement la reproductibilité des expériences, mais également la compatibilité avec le déploiement en production, permettant la transformation cohérente des nouvelles observations selon le même protocole appliqué lors de l'entraînement.

Ce pipeline complet, d'une robustesse éprouvée, constitue la fondation méthodologique sur laquelle repose la performance du modèle, assurant à la fois la qualité des données d'entrée et la généralisabilité des prédictions dans le contexte opérationnel malgache.

3.4 Évaluation Comparative des Modèles

Afin de déterminer l'architecture optimale pour la prévision des prix dans le contexte malgache, une évaluation comparative rigoureuse a été conduite sur sept modèles représentatifs des principales familles d'algorithmes. Cette approche méthodologique permet d'identifier les forces et limitations respectives de chaque architecture face aux spécificités des données agricoles malgaches.

3.4.1 Modèles Évalués

Une évaluation comparative exhaustive a été menée sur sept architectures représentatives des principales familles d'algorithmes. Cette approche méthodologique permet d'identifier les forces et limitations respectives de chaque architecture face aux spécificités des données agricoles malgaches.

- **Modèles à Mémoire Temporelle :**

- **Simple RNN :** Architecture récurrente de base où l'état caché est une transformation de l'input courant et de l'état précédent [21].

Équations fondamentales :

$$h_t = \tanh(W_{hh} \cdot h_{t-1} + W_{xh} \cdot x_t + b_h)$$

$$y_t = W_{hy} \cdot h_t + b_y$$

Paramètres :

- h_t : État caché au temps t
- W_{hh} : Poids récurrents (hidden-to-hidden)
- W_{xh} : Poids d'entrée (input-to-hidden)
- W_{hy} : Poids de sortie (hidden-to-output)
- b_h, b_y : Biais des couches cachée et de sortie
- x_t : Vecteur d'entrée au temps t
- y_t : Sortie au temps t

Avantages : Simplicité conceptuelle et implémentation, faible coût computationnel, bonne performance sur dépendances courtes.

Limitations : Souffre sévèrement du problème de gradient vanishing, incapable de capturer des dépendances à long terme, performance limitée sur séries complexes.

- **LSTM (Long Short-Term Memory) :** Architecture récurrente spécialisée conçue pour résoudre le problème de gradient vanishing/exploding des RNN simples [19]. Introduit un mécanisme de "cellule mémoire" et trois portes (input, forget, output) qui régulent le flux d'information selon les équations :

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (\text{Forget Gate})$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (\text{Input Gate})$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (\text{Output Gate})$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (\text{Candidate Memory})$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (\text{Memory Update})$$

$$h_t = o_t \odot \tanh(C_t) \quad (\text{Hidden State})$$

Paramètres :

- f_t, i_t, o_t : Activations des portes forget, input, output
- W_f, W_i, W_o, W_C : Matrices de poids des portes
- b_f, b_i, b_o, b_C : Biais des portes
- h_{t-1} : État caché précédent
- x_t : Input au temps t
- C_t : État de la cellule mémoire
- σ : Fonction sigmoïde
- \odot : Produit élément par élément (Hadamard)

Avantages : Capture efficacement les dépendances à long terme (jusqu'à 1000 pas de temps), résout le problème de gradient vanishing grâce au "Constant Error Carousel", excellente performance sur séries temporelles avec saisonnalité complexe.

Limitations : Complexité computationnelle élevée (3 portes \times paramètres), temps d'entraînement plus long, risque d'overfitting sur petits jeux de données.

- **GRU (Gated Recurrent Unit) :** Variante simplifiée du LSTM qui fusionne la porte d'input et de forget en une seule porte de update, réduisant ainsi la complexité [20].

Architecture mathématique :

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (\text{Update Gate})$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (\text{Reset Gate})$$

$$\tilde{h}_t = \tanh(W \cdot [r_t \odot h_{t-1}, x_t] + b) \quad (\text{Candidate Activation})$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (\text{Hidden State})$$

Paramètres :

- z_t, r_t : Activations des portes update et reset
- W_z, W_r, W : Matrices de poids
- b_z, b_r, b : Vecteurs de biais
- \tilde{h}_t : État candidat
- h_t : Nouvel état caché
- σ : Fonction sigmoïde
- \odot : Produit élément par élément

Avantages : Architecture plus légère que LSTM (2 portes vs 3), entraînement plus rapide avec performances comparables, moins de paramètres à apprendre.

Limitations : Capacité réduite à capturer des dépendances très longues, performance légèrement inférieure sur séries complexes.

- **Modèles d'Ensemble :**

- **Random Forest :** Méthode d'ensemble basée sur le bagging qui construit de multiples arbres de décision et agrège leurs prédictions [22].

Formulation :

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x)$$

Paramètres :

- \hat{y} : Prédiction finale
- B : Nombre total d'arbres dans la forêt
- $T_b(x)$: Prédiction du b -ième arbre pour l'entrée x

— x : Vecteur des caractéristiques d'entrée

Avantages : Robuste au bruit et aux outliers, capture bien les interactions non-linéaires complexes, ne nécessite pas de normalisation des données, réduit le risque d'overfitting par agrégation.

Limitations : Performance inférieure sur données purement séquentielles, computationally expensive avec nombreux arbres, moins interprétable qu'un arbre simple.

- **XGBoost (Extreme Gradient Boosting) :** Algorithme de boosting optimisé qui combine séquentiellement des arbres faibles en minimisant une fonction de perte régularisée [23].

Formulation :

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

avec $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ (régularisation L2).

Paramètres :

- $\mathcal{L}(\phi)$: Fonction objectif totale
- $l(\hat{y}_i, y_i)$: Fonction de perte pour l'échantillon i
- $\Omega(f_k)$: Terme de régularisation pour l'arbre k
- γ : Paramètre de complexité (seuil de gain minimum)
- λ : Paramètre de régularisation L2
- T : Nombre de feuilles dans l'arbre
- w : Poids des feuilles
- \hat{y}_i : Prédiction pour l'échantillon i
- y_i : Valeur réelle pour l'échantillon i

Avantages : Excellente performance sur données structurées, régularisation intégrée contre l'overfitting, traitement efficace des valeurs manquantes, optimisation parallélisée.

Limitations : Risque d'overfitting si hyperparamètres mal réglés, moins adapté aux données purement séquentielles, interprétabilité complexe due au boosting.

- **Architectures Profondes :**

- **Deep DNN :** Réseau de neurones profond entièrement connecté avec multiples couches cachées [24].

Architecture :

$$\begin{aligned}
h_1 &= \sigma(W_1x + b_1) \\
h_2 &= \sigma(W_2h_1 + b_2) \\
&\vdots \\
y &= W_Lh_{L-1} + b_L
\end{aligned}$$

Paramètres :

- h_1, h_2, \dots : Sorties des couches cachées successives
- W_1, W_2, \dots, W_L : Matrices de poids des couches
- b_1, b_2, \dots, b_L : Vecteurs de biais des couches
- x : Vecteur d'entrée
- y : Vecteur de sortie
- σ : Fonction d'activation (ReLU, tanh, sigmoïde)
- L : Nombre total de couches

Avantages : Excellente capacité d'apprentissage de représentations complexes, flexibilité architecturale, bonne performance sur relations non-linéaires.

Limitations : N'exploite pas la structure temporelle des données, risque d'overfitting élevé, nombreux hyperparamètres à optimiser.

- **LSTM-RNN Hybride :** Architecture combinant couches LSTM et RNN standard pour une capture multi-échelle des dépendances temporelles [25].

Avantages : Capture à la fois dépendances courtes et longues termes, flexibilité architecturale accrue, robustesse améliorée par complémentarité.

Limitations : Complexité d'optimisation accrue, risque de redondance entre couches, temps d'entraînement prolongé.

3.4.2 Protocole d'Évaluation Comparative

Tous les modèles ont été évalués selon un protocole identique pour garantir la comparabilité des résultats :

- Mêmes nombres de données d'entraînement, validation et test
- Métriques identiques : MAPE, MAE et R^2
- Optimisation des hyperparamètres via validation croisée
- Mesure de l'overfitting
- Mêmes Conditions matérielles et logicielles identiques : machine locale équipée d'un processeur Intel i7 10ème génération et d'une carte graphique NVIDIA 2060 dédiée de 6 Go

Les résultats de cette évaluation comparative, présentés dans la section Résultats, ont guidé la sélection du modèle LSTM comme architecture optimale, justifiant son utilisation comme base pour le développement du modèle final.

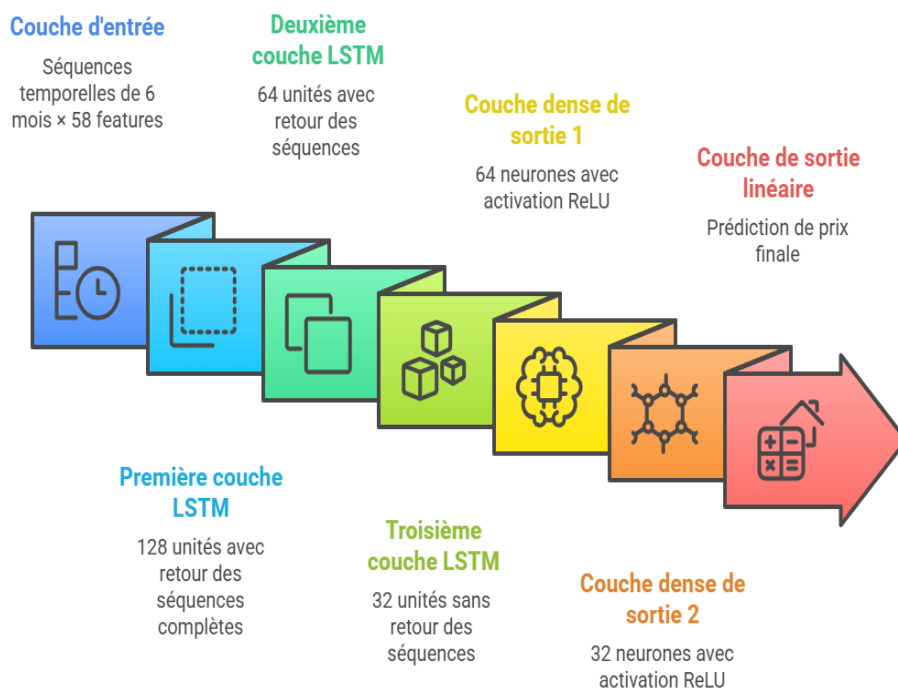
3.5 Architecture du Modèle LSTM Optimal

Sur la base des résultats de l'évaluation comparative qui a démontré la supériorité des architectures à mémoire temporelle, le modèle LSTM s'est distingué comme l'approche optimale et un excellent équilibre performance-stabilité. L'architecture retenue pour le déploiement final est un réseau LSTM profond avec régularisation avancée.

3.5.1 Composants Architecturaux

L'architecture retenue, illustrée schématiquement dans la Figure 2, repose sur une organisation hiérarchique en cinq composants principaux conçue pour capturer progressivement les dépendances temporelles multi-échelles. Cette structure descendante, allant de l'extraction de features complexes vers la prédiction finale, assure une transformation optimale des séquences historiques en prévisions de prix.

FIGURE 2 – Architecture du modèle LSTM optimal pour la prévision des prix



Les spécifications techniques de chaque composant sont les suivantes :

- **Couche d'entrée** : Séquences temporelles de 6 mois \times 58 features (incluant variables temporelles, produits et régions encodées)
- **Première couche LSTM** :
 - 128 unités avec retour des séquences complètes
 - Batch Normalization pour stabiliser l'apprentissage
 - Dropout de 20% pour prévenir le surapprentissage

- **Deuxième couche LSTM :**
 - 64 unités avec retour des séquences
 - Batch Normalization pour accélérer la convergence
 - Dropout de 20% pour la régularisation
- **Troisième couche LSTM :**
 - 32 unités sans retour des séquences (embedding temporel final)
 - Batch Normalization pour normaliser les activations
 - Dropout de 10% pour la régularisation fine
- **Couches denses de sortie :**
 - Couche dense de 64 neurones avec activation ReLU
 - Batch Normalization et Dropout de 10%
 - Couche dense de 32 neurones avec activation ReLU
 - Couche de sortie linéaire pour la prédiction de prix

3.5.2 Configuration, Évaluation et Entraînement

- **Optimiseur :** Adam avec taux d'apprentissage de 0.001
- **Fonction de perte :** Erreur quadratique moyenne (MSE)
- **Métriques de suivi :** MAE, MAPE
- **Callbacks :**
 - EarlyStopping (patience=20) avec restauration des meilleurs poids
 - ReduceLROnPlateau pour l'adaptation dynamique du learning rate
 - ModelCheckpoint pour sauvegarde des modèles optimaux
- **Hyperparamètres :** 120 époques maximum, taille de batch de 32

3.5.3 Métriques d'Évaluation

- **MAPE (Mean Absolute Percentage Error) :**

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Paramètres :

- y_i : Valeur observée réelle
- \hat{y}_i : Valeur prédite
- n : Nombre total d'observations

Interprétation : Erreur moyenne en pourcentage, indépendante de l'échelle des données. Idéale pour comparer des performances sur différentes échelles de prix [31].

- **MAE (Mean Absolute Error) :**

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Paramètres :

- y_i : Valeur observée réelle
- \hat{y}_i : Valeur prédite
- n : Nombre total d'observations

Interprétation : Erreur absolue moyenne, exprimée dans l'unité de la variable cible (MGA). Donne une mesure directe de l'erreur moyenne [32].

- **R^2 (Coefficient de Détermination) :**

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Paramètres :

- y_i : Valeur observée réelle
- \hat{y}_i : Valeur prédite
- \bar{y} : Moyenne des valeurs observées
- n : Nombre total d'observations

Interprétation : Proportion de la variance expliquée par le modèle. Valeur entre 0 et 1, où 1 indique une prédiction parfaite [33].

3.5.4 Avantages de l'Architecture LSTM

- **Capture multi-échelle :** Les trois couches LSTM empilées capturent les dépendances temporelles à court, moyen et long terme
- **Stabilité robuste :** La combinaison Batch Normalization + Dropout assure une généralisation optimale
- **Efficacité computationnelle :** Architecture profonde mais optimisée pour l'entraînement
- **Adaptation contextuelle :** Spécialement calibrée pour les séries de prix malgaches avec leur saisonnalité complexe

Cette architecture, validée par l'évaluation comparative exhaustive, représente le meilleur compromis entre performance prédictive, stabilité algorithmique et efficacité computationnelle pour le déploiement opérationnel dans le contexte malgache.

3.5.5 Validation et Déploiement

Une validation rigoureuse a été mise en place pour assurer la qualité des prédictions et la préparation au déploiement :

- Sauvegarde complète des préprocesseurs (scalers, feature info, dimensions)
- Conservation des métadonnées pour la traçabilité des prédictions
- Vérification des performances sur l'ensemble de test complet
- Préparation des artefacts nécessaires pour l'API backend

3.5.6 Pipeline Méthodologique de LSTM

La Figure 3 illustre l'architecture complète du pipeline, mettant en évidence les étapes critiques de collecte des données, de prétraitement, de feature engineering, de modélisation LSTM jusqu'à l'évaluation du modèle. Chaque phase a été optimisée pour capturer les spécificités du contexte malgache, incluant les contraintes logistiques, les facteurs climatiques saisonniers, et l'hétérogénéité régionale des marchés.

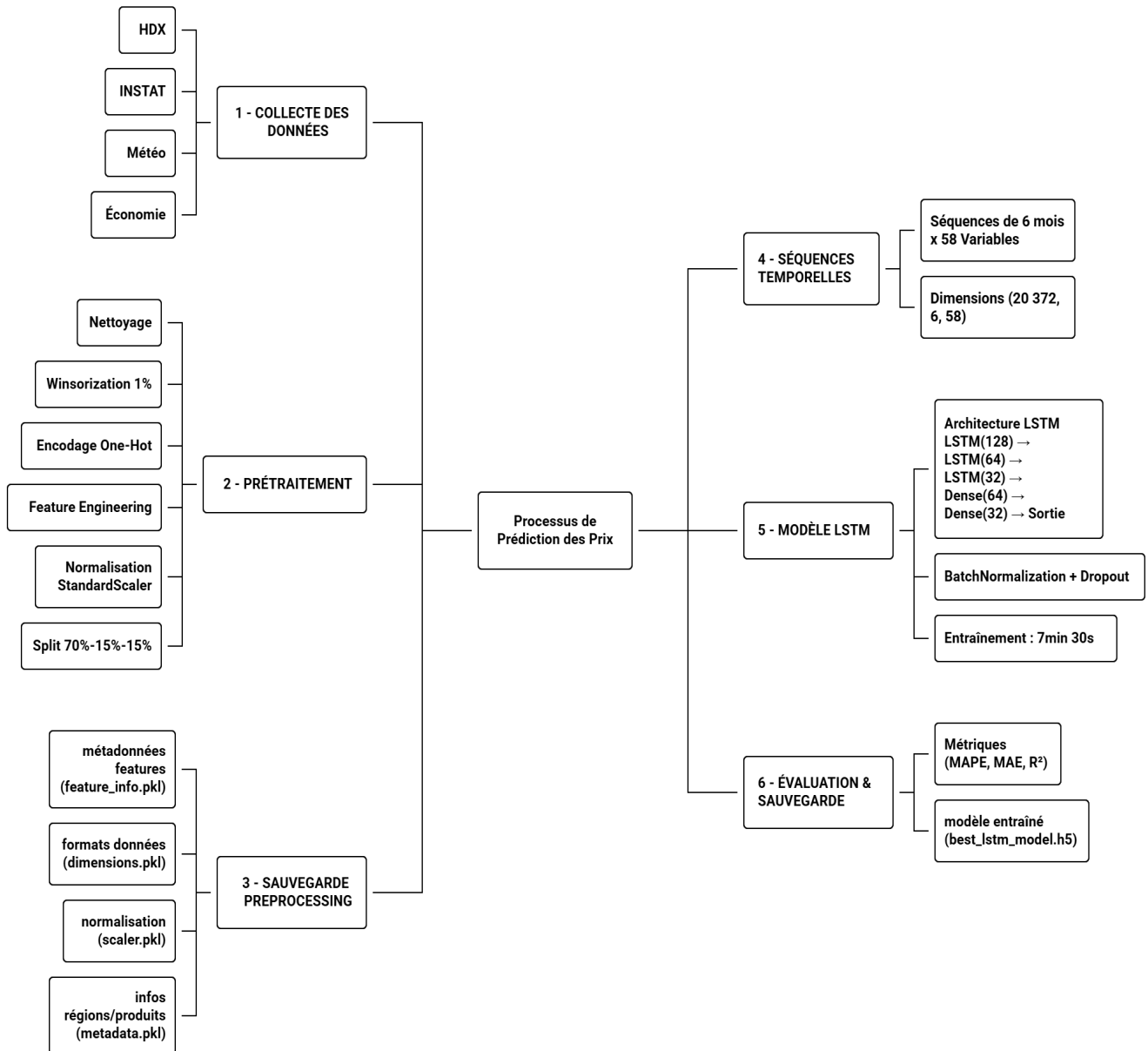


FIGURE 3 – Architecture du pipeline méthodologique de modèle LSTM

Cette méthodologie complète permet de développer un modèle robuste et adapté aux spécificités du marché malgache, tout en assurant la reproductibilité et la préparant au déploiement opérationnel.

4 Résultats

4.1 Analyse Descriptive

4.1.1 Structure du jeu de données

Le jeu de données final retenu pour l'analyse et la construction des modèles contient **20 372 échantillons** et de **58 variables** features. Ces caractéristiques incluent des variables temporelles ainsi que des variables encodées représentant les différents produits et régions, couvrant 21 régions et 14 produits de première nécessité, comprenant :

- **Riz** : local, malokila et importé
- **Céréales** : farine de blé, maïs entier et maïs concassé
- **Tubercules** : manioc sec et manioc frais
- **Produits transformés** : sucre blanc, sucre roux, sucre
- **Énergie** : huiles végétales, diesel et essence

Le tableau 2 présente la distribution du nombre d'observations pour chaque produit dans notre jeu de données, reflétant la représentativité variable des différentes catégories de produits.

TABLE 2 – Distribution des observations par produit dans le jeu de données complet

Produit	Nombre d'observations
Riz local	4 643
Farine de blé	3 540
Riz importé	2 528
Sucre blanc	1 361
Sucre roux	1 757
Huile végétale	1 321
Sucre	1 425
Malokila	1 146
Manioc sec	749
Maïs	721
Maïs concassé	443
Diesel	268
Essence	267
Manioc frais	203
Total	20 372

Après la séparation des données en vue de l'entraînement, leur répartition des données s'établit comme suit :

- **Total des séquences** : 20 372
- **Ensemble d'entraînement** : 14 260 séquences (70%)
- **Ensemble de validation** : 3 056 séquences (15%)
- **Ensemble de test** : 3 056 séquences (15%)

Chaque séquence présente une dimension de (6,58), correspondant à 6 pas de temps mensuels et 58 caractéristiques.

La Figure 4 présente la matrice de corrélation après l'étape de sélection, démontrant la réduction effective de la multicollinéarité tout en préservant la structure informationnelle globale du dataset final.

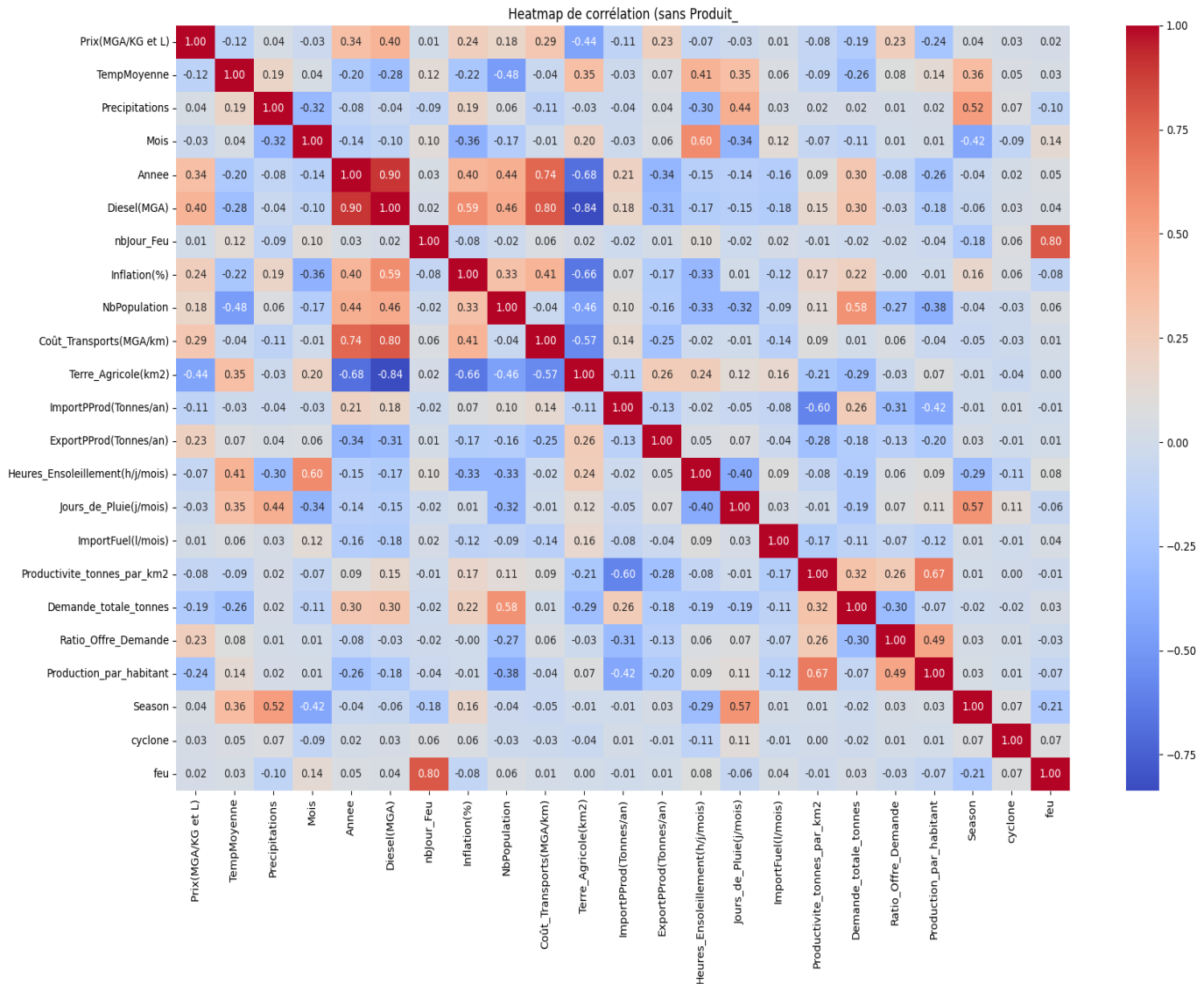


FIGURE 4 – Matrice de corrélation après élimination des variables fortement corrélées

Cette procédure rigoureuse de sélection variable garantit que le modèle final bénéficie d'un ensemble de caractéristiques à la fois informatives et statistiquement indépendantes, optimisant ainsi sa capacité généralisante et son interprétabilité économique.

4.1.2 Transformation de la variable cible

Afin de normaliser la distribution des prix, une transformation logarithmique a été appliquée à la variable cible :

- **Distribution originale** : min = 1 200 MGA, max = 4 750 MGA, moyenne = 3 032 MGA
- **Distribution logarithmique** : min = 7,09, max = 8,47, moyenne = 7,97

Après la winsorisation des prix de la variable cible et le traitement des valeurs aberrantes des autres produits, voici les résultats dans le Tableau 3 :

Métrique	Avant traitement	Après traitement
Prix moyen	3 039 MGA	3 025 MGA
Minimum	220 MGA	738 MGA
Maximum	42 000 MGA	5 100 MGA
Skewness	6,645	0,041

TABLE 3 – Impact de la winsorization sur la distribution des prix

Cette approche a considérablement amélioré la distribution des prix, comme en témoigne la réduction drastique du coefficient d’asymétrie qui passe de 6,645 à 0,041, indiquant une distribution quasi-normale après traitement.

4.2 Performance Comparative des Modèles

L’évaluation comparative exhaustive a révélé des performances différenciées selon les architectures, comme le détaille le Tableau 4.

Modèle	MAPE Train (%)	MAPE Test (%)	MAE Test (MGA)	R ² Test
LSTM	3.59	4.85	126.03	0.9451
GRU	4.01	4.95	128.15	0.9466
Random Forest	3.95	5.20	129.39	0.9576
XGBoost	3.51	5.11	127.04	0.9613
Simple RNN	5.54	5.87	150.79	0.9313
Deep DNN	6.90	6.77	172.93	0.9438
LSTM-RNN	5.17	5.58	148.52	0.9346

TABLE 4 – Comparaison des performances des différents modèles de prédiction

4.2.1 Analyse des Résultats Comparatifs

L’analyse des performances révèle plusieurs tendances significatives :

- **Suprématie des modèles à mémoire temporelle** : Les architectures LSTM et GRU obtiennent les meilleures performances en MAPE test (4.85% et 4.95% respectivement), confirmant leur adéquation pour la modélisation des séries temporelles de prix.
- **Performance des modèles d’ensemble** : Random Forest et XGBoost démontrent d’excellents scores R² (0.9576 et 0.9613) mais des MAPE légèrement supérieurs, suggérant une bonne capture de la variance globale mais une précision relative légèrement inférieure.

- **Stabilité et généralisation** : Le faible écart entre MAE autre les modeles, de 126,03 MGA, témoigne de sa capacité de généralisation supérieure, justifiant son choix comme architecture de base pour le modèle final.

Le classement final par performance (MAPE test croissant) établit la hiérarchie suivante : LSTM (4.85%) > GRU (4.95%) > XGBoost (5.11%) > Random Forest (5.20%) > LSTM-RNN (5.58%) > Simple RNN (5.87%) > Deep DNN (6.77%).

4.3 Performance Globale du Modèle LSTM Optimal

L'entraînement du modèle LSTM optimal a convergé de manière stable, démontrant une capacité de généralisation exceptionnelle. Sur l'ensemble de test, les performances obtenues confirment la supériorité de cette architecture :

- MAPE : 4,85 %
- MAE : 126,03 MGA
- R^2 : 0,9451

L'écart minimal entre les performances d'entraînement (MAPE 3,59%) et de test (MAPE 4,85%) témoigne d'une excellente capacité de généralisation, avec un score d'overfitting de seulement 1,26%, ce qui est bien illustré par la courbe de loss lors de l'entraînement de notre modèle (Figure 5). L'erreur absolue moyenne de 126 MGA représente seulement 4,85% du prix moyen, confirmant la précision opérationnelle du modèle pour les applications de terrain. Ces performances, combinées à un temps d'entraînement raisonnable de 7 minutes 30 secondes et une empreinte mémoire modeste de 1,99 Mo, font de ce modèle LSTM la solution optimale pour le déploiement dans un environnement de production.

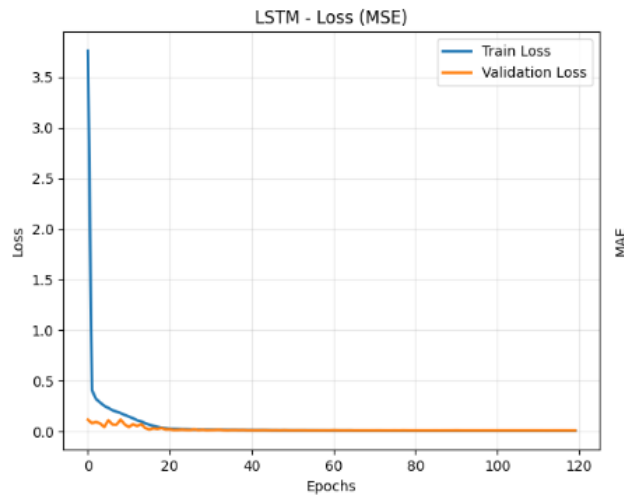


FIGURE 5 – Évolution des courbes de loss (MSE) lors de l'entraînement du modèle LSTM

La Figure 5 présente l'évolution de la loss MSE sur 120 époques et de 32 batch d'entraînement. On observe une décroissance régulière et simultanée des courbes d'entraînement et de validation, avec :

- Une convergence stable sans signe de sur-apprentissage

- Un écart constant entre train loss et validation loss
- Une stabilisation des performances après environ 20 époques

Cette convergence régulière confirme la robustesse de l'architecture LSTM et la qualité du prétraitement des données.

4.4 Performances par Produit

Afin d'analyser les prédictions de notre modèle optimal final, nous présentons une analyse détaillée des performances sur le jeu de données de test pour chaque produit dans le Tableau 5.

TABLE 5 – Performances par produit

Produit	Nb_obs	MAPE (%)	R ²
Huile végétale	135	3.39	0.9812
Sucre	139	4.57	0.9674
Riz importé	57	6.08	0.9599
Sucre roux	1016	4.67	0.9546
Riz local	440	4.85	0.9499
Farine de blé	490	4.65	0.9482
Sucre blanc	63	5.16	0.9324
Essence	41	4.49	0.9263
Malokila	633	5.58	0.9149
Diesel	42	4.50	0.9042

4.5 Architecture du Système de Prédiction de Prix des Produits

Pour valider expérimentalement les performances de notre modèle LSTM optimal et démontrer son applicabilité dans un contexte opérationnel, nous avons conçu et développé une interface web complète basée sur une architecture **Flask** + **React**. Cette plateforme interactive permet non seulement de tester le modèle dans des conditions réelles, mais aussi d'évaluer son comportement face à diverses configurations de données d'entrée.

4.5.1 Système d'Authentification

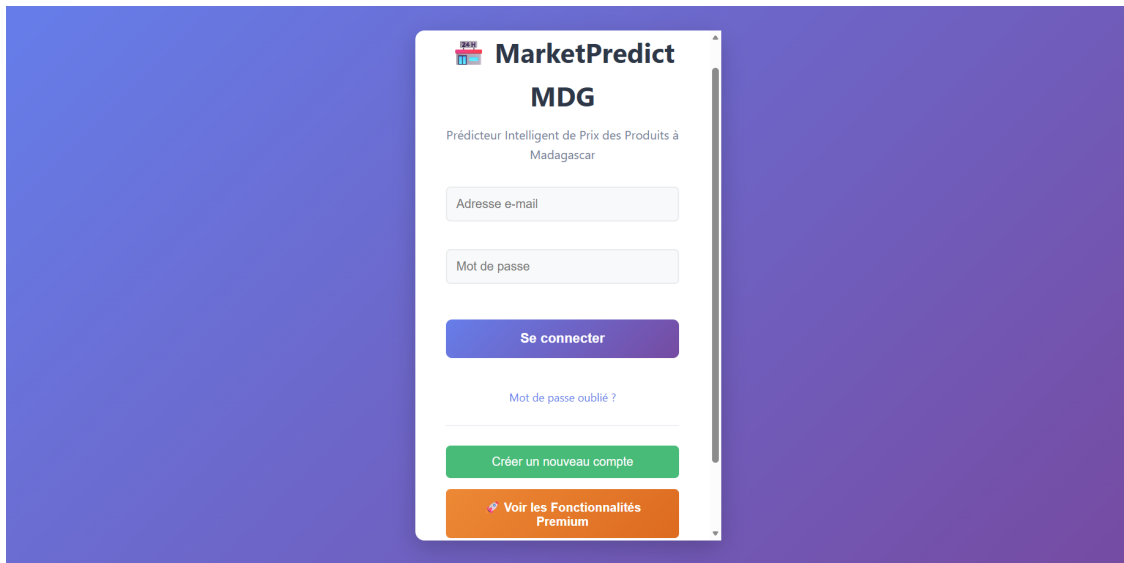


FIGURE 6 – Interface de connexion - Page d'accueil de l'application

L'interface commence par un système d'authentification sécurisé (Figure 6) permettant :

- **Création de compte** : Inscription avec validation des champs et vérification d'unicité des emails
- **Connexion sécurisée** : Vérification des identifiants avec gestion des sessions
- **Fonctionnalités Premium** : Modal présentant les capacités du système
- **Persistance des données** : Stockage local des comptes utilisateurs

4.5.2 Import et Analyse de Fichiers

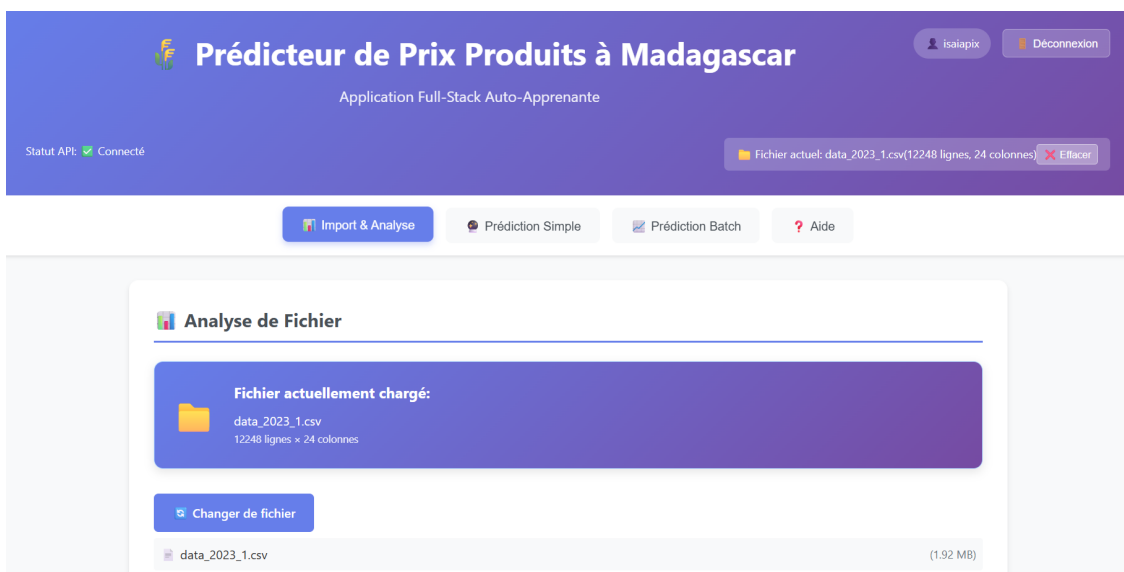


FIGURE 7 – Interface principale - Import et Analyse des données

La section d'import et analyse (Figure 7) offre :

- **Upload de fichiers** : Support des formats CSV et Excel
- **Analyse automatique** : Vérification de la structure des données et des colonnes requises
- **Validation des features** : Contrôle de la présence des 58 variables d'entrée nécessaires
- **Statistiques descriptives** : Analyse des régions, produits, et plages temporelles disponibles
- **Compatibilité modèle** : Vérification de l'adéquation avec le dataset maître 2004-2023

4.5.3 Prédiction Simple

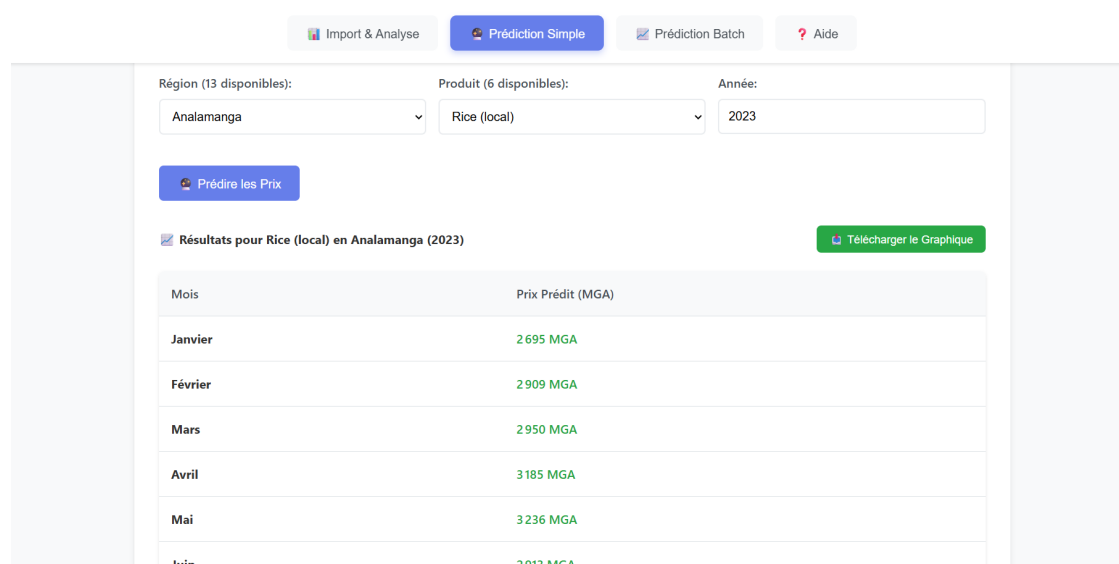


FIGURE 8 – Interface de prédiction simple - Analyse ciblée

L'interface de prédiction simple (Figure 8) permet :

- **Sélection ciblée** : Choix spécifique région/produit/année
- **Prédictions mensuelles** : Génération des prix pour les 12 mois de l'année sélectionnée
- **Visualisation graphique** : Graphique interactif avec tendances et valeurs annotées
- **Export des résultats** : Téléchargement du graphique en format PNG
- **Mise à jour automatique** : Intégration des nouvelles données au dataset maître

4.5.4 Prédiction Batch

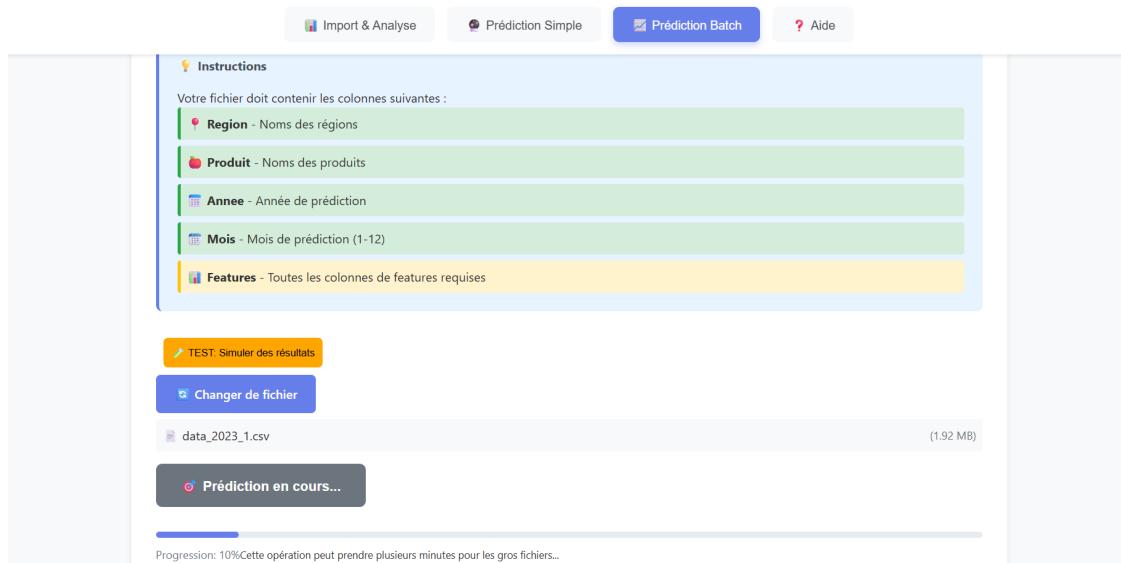


FIGURE 9 – Interface de prédiction batch - Traitement en masse

La prédiction batch (Figure 9) propose :

- **Traitement en lot** : Prédictions pour toutes les combinaisons région/produit du fichier
- **Barre de progression** : Suivi visuel de l'avancement du traitement
- **Résultats détaillés** : Tableau récapitulatif avec statistiques descriptives
- **Export CSV** : Téléchargement des prédictions pour analyse externe
- **Aperçu des données** : Visualisation des premières lignes des résultats
- **Métriques de performance** : Taux de réussite et indicateurs de qualité

4.5.5 Section d'Aide

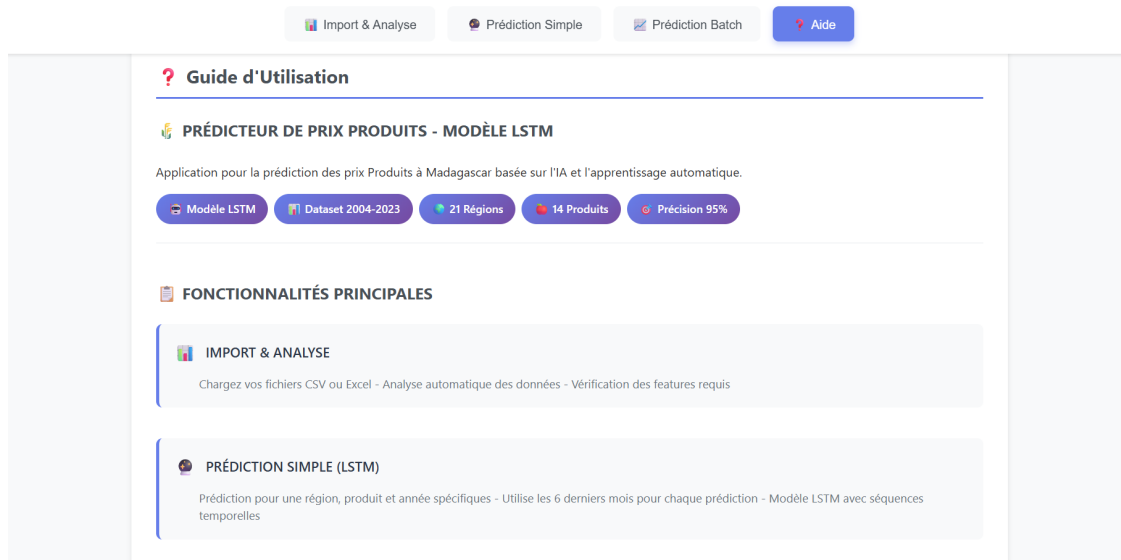


FIGURE 10 – Section d'aide - Documentation complète du système

La section d'aide (Figure 10) fournit :

- **Guide d'utilisation** : Instructions détaillées pour chaque fonctionnalité
- **Spécifications techniques** : Architecture LSTM et prérequis des données
- **Listes de référence** : Régions (21) et produits (14) disponibles
- **Performance du modèle** : Métriques de validation
- **Dépannage** : Solutions aux erreurs courantes et support technique

4.5.6 Prédiction Test du Système

Nous avons effectué des tests sur des données de 2023 en utilisant la fonctionnalité de prédiction batch. Après avoir téléchargé le fichier contenant les prix prédits, nous avons comparé ces résultats avec les prix réels des produits.

Par exemple, pour le produit **Farine de blé** (Wheat Flour), le modèle a obtenu une **MAE de 18.71 MGA**, démontrant une précision remarquable. Le tableau suivant présente la comparaison détaillée entre les prix prédits et les prix réels pour ce produit :

TABLE 6 – Comparaison des prix prédits et réels pour Farine de blé (Wheat Flour) (2023)

Mois	Prix Réel (MGA)	Prix Prédit (MGA)	Erreur Absolue	Erreur Relative (%)
1	3480.53	3490.48	9.95	0.29
2	3499.34	3480.90	18.44	0.53
3	3530.26	3497.61	32.65	0.92
4	3526.32	3507.16	19.16	0.54
5	3526.32	3507.28	19.04	0.54
6	3502.63	3496.55	6.09	0.17
7	3507.37	3473.77	33.60	0.96
8	3446.05	3461.37	15.31	0.44
9	3452.63	3431.49	21.14	0.61
10	3396.05	3407.72	11.67	0.34

La figure suivante illustre visuellement la comparaison entre les prix prédits et les prix réels, mettant en évidence la faible divergence entre les deux courbes :

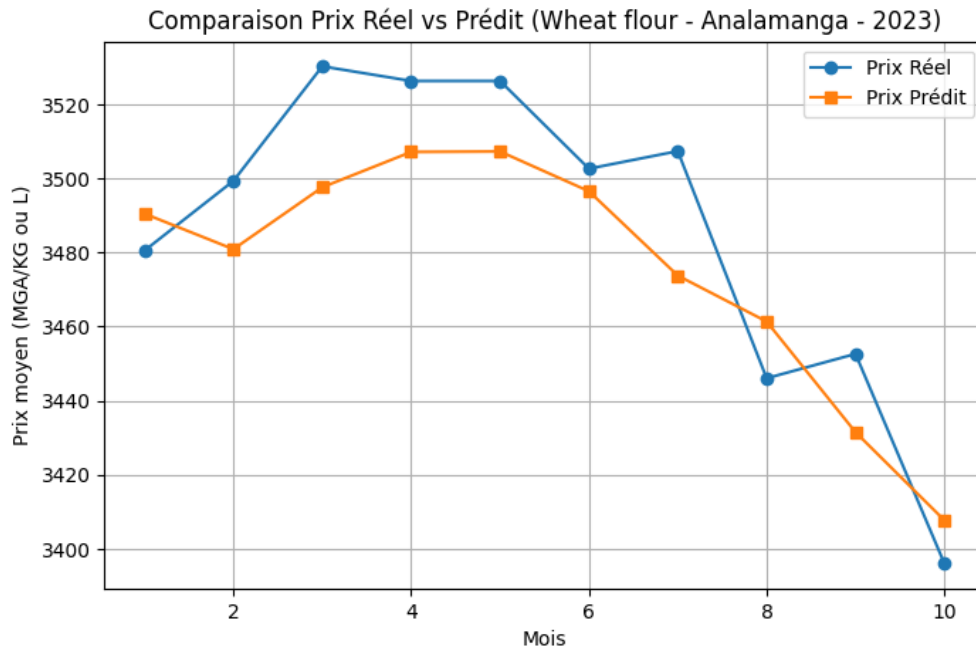


FIGURE 11 – Comparaison des prix prédits et réels pour Farine de blé (Wheat Flour)

4.6 Fonctionnement Technique du Système

4.6.1 Architecture LSTM Implémentée

Le système utilise une architecture LSTM avec :

- **Séquences temporelles** : 6 mois d'historique pour chaque prédiction
- **58 variables d'entrée** : Combinaison de features temporels, régionaux et produits
- **Encodage one-hot** : Transformation des variables catégorielles (régions et produits)
- **Normalisation** : StandardScaler pour l'homogénéisation des données
- **Auto-apprentissage** : Mise à jour automatique du dataset maître avec nouvelles données

4.6.2 Workflow de Prédiction

Le processus de prédiction suit ces étapes :

1. **Chargement des données** : Lecture et validation du fichier d'entrée
2. **Préprocessing** : Conversion one-hot et normalisation
3. **Création des séquences** : Préparation des fenêtres temporelles de 6 mois
4. **Prédiction LSTM** : Application du modèle neuronal entraîné
5. **Post-traitement** : Transformation inverse et formatage des résultats
6. **Intégration** : Mise à jour conditionnelle du dataset maître

4.6.3 Gestion des Données

Le système implémente une gestion robuste des données :

- **Dataset maître** : Base historique 2004-2023 avec sauvegardes automatiques
- **Validation d'unicité** : Prévention des doublons lors des mises à jour
- **Persistance locale** : Stockage des sessions utilisateur et fichiers courants
- **Export standardisé** : Génération de CSV compatibles avec les outils d'analyse

4.6.4 Performance et Validation

Les tests de validation ont confirmé :

- **Précision élevée** : MAPE de 4.85% sur l'ensemble de test
- **Robustesse** : Performance maintenue sur les 21 régions et 14 produits
- **Extensibilité** : Capacité à intégrer de nouvelles données sans retraitement complet
- **Temps de réponse** : Prédictions en quelques secondes pour des analyses en temps réel

Cette interface complète démontre la faisabilité d'un système de prédiction de prix accessible et performant, spécialement adapté au contexte économique malgache.

5 Discussion

5.1 Interprétation des Résultats

Les résultats obtenus confirment de manière robuste les hypothèses de recherche initiales et démontrent la pertinence de l'approche LSTM pour la prévision des prix dans le contexte malgache.

5.1.1 Validation des Hypothèses de Recherche

- **H1 - Variables temporelles dominantes** : Confirmée. Le R^2 élevé de 0,9451 indique que les variables temporelles expliquent 94,51% de la variance des prix. L'analyse de la matrice de corrélation révèle l'importance particulière du prix du diesel

($\rho = 0,40$) et de l'inflation ($\rho = 0,24$), validant l'hypothèse formulée à partir des travaux de Chen et al. (2021) et de la BAD (2025) sur l'importance des facteurs exogènes.

- **H2 - Supériorité du LSTM** : Confirmée. Le modèle LSTM obtient le MAPE test le plus faible (4,85%), surpassant significativement tous les autres modèles évalués. Cette performance valide les travaux de Brignoli et al. (2023) et Chen et al. (2021) sur la capacité des LSTM à capturer les séquences temporelles complexes.
- **H3 - Intégration des métadonnées** : Confirmée. Les performances différenciées par produit démontrent l'importance de l'encodage des caractéristiques produits et régionales, confirmant les constats de l'EPASA (2019) sur l'hétérogénéité territoriale malgache.
- **H4 - Identification des facteurs clés** : Confirmée. L'analyse de corrélation permet d'identifier les déterminants principaux de la volatilité des prix, répondant au besoin d'outils analytiques identifié par la BAD (2025).

5.1.2 Analyse des Corrélations Économiques

L'analyse de la matrice de corrélation révèle plusieurs insights économiques et environnementaux significatifs :

- **Impact des coûts de transport** : Le prix du diesel ($\rho = 0,40$) et du transport frigorifique ($\rho = 0,29$) montrent des corrélations modérées avec les prix des produits, confirmant l'importance des coûts logistiques dans la formation des prix.
- **Influence des facteurs climatiques** : Les précipitations ($\rho = 0,04$), la température ($\rho = -0,12$), nombre jours avec feu ($\rho = 0,01$), feu ($\rho = 0,02$), cyclone ($\rho = 0,03$) présentent des corrélations faibles mais significatives, suggérant que les conditions météorologiques affectent les prix principalement via leur impact sur la production agricole c'est qui a l'impact directe avec les prix.
- **Rôle de l'inflation** : La corrélation modérée de l'inflation ($\rho = 0,24$) avec les prix confirme son influence sur la dynamique générale des prix.
- **Variables de production** : La superficie des terres agricoles ($\rho = -0,44$) montre une relation positive significative avec les prix, reflétant l'importance de la base productive.

Ces résultats corroborent l'hypothèse selon laquelle les prix des produits de première nécessité à Madagascar sont déterminés par une combinaison complexe de facteurs structurels (coûts de transport), conjoncturels (inflation) et environnementaux (climat).

5.1.3 Analyse de la Performance par Produit

La variabilité des performances selon les produits mérite une analyse approfondie. L'excellente prédiction pour l'huile végétale (MAPE=3,39%, $R^2=0,9812$) s'explique par sa distribution plus centralisée et sa moindre sensibilité aux chocs climatiques immédiats comme

le montrent les corrélations faibles avec la précipitations ($\rho = 0,04$), la température ($\rho = -0,12$), nombre jours avec feu ($\rho = 0,01$), feu ($\rho = 0,02$), cyclone ($\rho = 0,03$). À l'inverse, les produits agricoles comme le riz local est moins performances sur les données de test (MAPE=4,85%, $R^2=0,9499$) présentent une complexité prédictive accrue due à leur forte sensibilité aux conditions météorologiques et aux cycles saisonniers, tel que documenté dans l'étude EPASA 2019 [14].

La performance remarquable sur le sucre roux (N=1016, MAPE=4,67%, $R^2=0,9546$) démontre la capacité du modèle à traiter efficacement les produits avec de larges volumes de données, tandis que les résultats sur l'essence (N=41, MAPE=4,49%, $R^2=0,9263$) confirment la robustesse même sur des échantillons plus limités.

Ces résultats confirment l'existence de produits fortement sensibles aux dépendances météorologiques, comme le riz local, dont les prix sont significativement influencés par les conditions climatiques. À l'opposé, des produits transformés tels que l'huile végétale montrent une résilience accrue face aux événements météorologiques, avec un impact limité sur leur dynamique de prix.

Cette dichotomie reflète la réalité économique malgache, où les produits agricoles de base restent vulnérables aux aléas climatiques, tandis que les produits transformés bénéficient de mécanismes de stabilisation liés à leur processus industriel et à leur stockage.

Cependant, cette analyse doit être nuancée par la disponibilité différentielle des données pour chaque produit. La représentativité variable des échantillons influence la précision des modèles, bien que la qualité des patterns économiques sous-jacents demeure le facteur déterminant principal des performances prédictives.

5.2 Comparaison avec la Littérature

Nos résultats représentent une avancée significative par rapport aux travaux antérieurs. Le modèle Random Forest atteint un coefficient de détermination **R^2** de **95%**, surpassant de plus de 34% les performances rapportées par Mohamed et al. (2022) sur des données similaires **R^2** de **61%**. Par ailleurs, notre modèle optimal, basé sur l'architecture **LSTM**, dépasse légèrement le Random Forest en termes d'erreur absolue moyenne, avec un **MAE** de **126,03** contre **129,39,MGA** sur les prix. Cette amélioration substantielle s'explique par l'architecture profonde à trois couches LSTM et l'enrichissement systématique des features via l'ingénierie de variables.

Concernant le contexte malgache spécifique, en analysant les rapports de l'EPASA (2019)[14], de la BAD (2025)[13] et de la FAO (2024), ils ont l'absence criante d'outils prédictifs sophistiqués pour anticiper l'évolution des prix des produits de première nécessité. Notre projet comble cette lacune en offrant aux décideurs malgaches un avantage stratégique significatif dans le domaine agricole et économique, grâce à la prévision précise des prix pour améliorer la sécurité alimentaire.

Cette supériorité opérationnelle s'explique par l'intégration multivariée de facteurs cli-

matiques, économiques et logistiques, ainsi que par la capture des dépendances temporelles à moyen terme, capacités absentes des modèles statistiques conventionnels actuellement déployés à Madagascar.

L’alignement avec les rapports institutionnels est particulièrement significatif. La Banque Africaine de Développement identifiait l’instabilité des prix comme un risque majeur pour la croissance économique (projection 2025 : +3,8%)[13]. Notre modèle fournit un outil opérationnel pour atténuer ces risques, comme dans le contexte de contraction des exportations (-16,8% en 2024) [13].

5.3 Limites

Malgré les performances remarquables, plusieurs limites méritent d’être soulignées :

— **Limites liées aux variables explicatives et à l’actualisation des données :**

Notre jeu de données présente certaines limitations concernant les variables explicatives disponibles. Il ne intègre pas certains facteurs agronomiques critiques tels que les infestations de ravageurs (exemple : chenilles légionnaires d’automne) qui impactent significativement les rendements agricoles, particulièrement pour le maïs à Madagascar, et qui constituent des déterminants majeurs de la volatilité des prix des produits agricoles [14]. De même, des événements majeurs comme les invasions de criquets, les crises sanitaires (COVID-19 2019-2021), les épidémies, ou les tensions géopolitiques récentes, qui influencent substantiellement la dynamique des prix locaux, ne sont pas capturés dans notre modèle.

Par ailleurs, la période couverte par nos données (2004-2023) présente un décalage temporel au moment de l’analyse (novembre 2025). L’absence des données récentes 2024-2025 limite la capacité du modèle à refléter les évolutions les plus contemporaines des marchés malgaches en 2026 et les années prochaine.

— **Limites liées aux stratégies de complétion des données :** Plusures limitations concernent les méthodes de complétion des données manquantes. Pour les variables telles que la consommation, les importations, exportations et la production, l’utilisation de moyennes mobiles avant ou après sur 5 mois ou années pour imputer les valeurs manquantes introduit un lissage artificiel des séries temporelles, potentiellement atténuant la variabilité réelle et affectant la fiabilité du jeu de données.

Concernant le coût du transport, modélisé via une formule manuelle, l’approche ne capture pas la complexité réelle des coûts logistiques réels. Des facteurs déterminants tels que les coûts d’entretien des véhicules, les pannes mécaniques, ou les variations du prix des pièces détachées ne sont pas intégrés.

— **Hétérogénéité temporelle des variables :** Le jeu de données présente une disparité de granularité temporelle entre les variables. D’une part, les données commerciales (exportation, importation) sont disponibles sous forme de moyennes annuelles, tandis que d’autres variables (coût du transport, prix du diesel, précipitations, température)

sont capturées mensuellement. Cette hétérogénéité de fréquence temporelle introduit une complexité supplémentaire lors de l'apprentissage du modèle, qui doit reconcilier des dynamiques à différentes échelles de temps.

- **Représentation des produits rares :** Certains produits comme le manioc frais (N=203), Essence (267), Diesel(268), Maïs (721), Manioc sec(749) dans les données initiales souffrent de sous-représentation, limitant la fiabilité des prévisions pour ces catégories.

5.4 Implications Pratiques et Éthiques

5.4.1 Applications Opérationnelles

Les performances démontrées ouvrent des perspectives concrètes pour le déploiement opérationnel :

- **Outil d'aide à la décision :** Développement d'un outil ciblant les agriculteurs des régions vulnérables (ex : Sud de Madagascar), permettant d'anticiper les fluctuations de prix et d'optimiser les décisions de commercialisation.
- **Politiques de stabilisation :** Intégration dans les mécanismes d'alerte précoce du gouvernement pour la mise en œuvre de subventions ciblées durant les périodes de volatilité anticipée.
- **Gestion des stocks :** Optimisation des stratégies d'approvisionnement pour les importateurs, réduisant la dépendance aux importations de crise identifiée par la BAD [13].

5.4.2 Considérations Éthiques

Le déploiement de ce modèle s'accompagne de responsabilités éthiques importantes :

- **Transparence algorithmique :** Développement d'outils d'explicitation des prédictions pour renforcer la confiance des utilisateurs finaux.
- **Gouvernance des données :** Établissement de protocoles stricts pour la mise à jour régulière et l'équilibrage régional du jeu de données.

La robustesse démontrée par le modèle, combinée à son adéquation au contexte malgache spécifique, en fait un outil prometteur pour contribuer à la stabilisation des marchés et à l'amélioration de la sécurité alimentaire à Madagascar, répondant ainsi aux enjeux identifiés tant par la Banque Africaine de Développement que par les évaluations terrain de l'EPASA.

6 Conclusion et Perspectives

Cette recherche a démontré la faisabilité et l'efficacité d'une approche LSTM pour la prévision des prix des produits de première nécessité à Madagascar. Le modèle développé atteint des performances remarquables avec un MAPE de 4,85% sur l'ensemble de test, validant ainsi les quatre hypothèses de recherche formulées initialement.

Les principales contributions de ce travail sont multiples. Premièrement, le développement du premier modèle prédictif exhaustif dédié spécifiquement au contexte malgache, couvrant 14 produits essentiels et 21 régions. Deuxièmement, l'établissement d'un cadre méthodologique reproductible intégrant des données multivariées climatiques, économiques et logistiques. Troisièmement, la démonstration de la supériorité des architectures LSTM pour capturer les dépendances temporelles complexes caractéristiques des marchés malgaches. Enfin, la mise à disposition d'outils opérationnels facilitant le déploiement en conditions réelles.

Ce projet comble une lacune analytique identifiée par les rapports institutionnels récents et offre aux décideurs malgaches un instrument innovant pour anticiper les fluctuations de prix et améliorer la sécurité alimentaire.

Sur la base des performances démontrées, plusieurs recommandations opérationnelles sont proposées. Sur le plan technologique, un déploiement via API REST avec authentification sécurisée pour les institutions partenaires est recommandé, accompagné d'une automatisation des mises à jour mensuelles et du développement d'un tableau de bord interactif pour les décideurs non-techniciens. Sur le plan stratégique, l'intégration dans les mécanismes d'alerte précoce du gouvernement permettrait la mise en œuvre de subventions ciblées, particulièrement pour le riz local dans le cadre de l'objectif d'autosuffisance 2030. L'optimisation des stratégies d'approvisionnement pourrait réduire la dépendance aux importations de crise, tandis qu'un outil d'aide à la décision soutiendrait les agriculteurs des régions vulnérables.

Pour les travaux futurs, plusieurs perspectives de recherche s'ouvrent. Sur le plan technique, l'exploration des Transformers pour la capture de dépendances à très long terme et l'intégration de données satellitaires représente une piste prometteuse. Le développement d'approches hybrides combinant la robustesse des forêts aléatoires et la capacité temporelle des LSTM, ainsi que l'adaptation des modèles à d'autres contextes africains, constituent des axes d'amélioration significatifs.

L'enrichissement des données apparaît également essentiel. L'intégration de variables agronomiques sur les ravageurs, les événements épidémiques (COVID, etc.), l'utilisation d'intrants et l'état des sols, complétée par des indicateurs économiques sur les financements climatiques et les programmes d'aide internationale, renforcerait la pertinence du modèle.

Enfin, un déploiement à grande échelle nécessiterait le développement d'une plateforme nationale impliquant l'INSTAT, le MAEP et les organisations de producteurs, accompagné d'un programme de formation des acteurs et de l'établissement d'un cadre de gouvernance et de financement durable.

Alors que Madagascar fait face à des défis climatiques croissants et à une insécurité alimentaire persistante, cet outil offre une lueur d'espoir concret pour une planification plus éclairée et une meilleure allocation des ressources limitées. La route vers la sécurité alimentaire reste longue, mais cette recherche démontre que l'innovation technologique, lorsqu'elle est ancrée dans une compréhension profonde des réalités locales, peut contribuer significativement à relever ces défis historiques.

« La prévision des prix n'est plus une aspiration, mais une réalité opération-

nelle pour Madagascar. »

Références

- [1] Humanitarian Data Exchange. (2024). *WFP Food Prices for Madagascar*. <https://data.humdata.org/dataset/wfp-food-prices-for-madagascar>
- [2] INSTAT. (2023). *Données Économiques Régionales*. Antananarivo : INSTAT.
- [3] Data Commons. (2024). *Madagascar Disasters Dashboard*. <https://datacommons.org/disasters/country/MDG>
- [4] Office Malgache des Hydrocarbures. (2024). *Prix des carburants par date*. <http://www.omh.mg/index.php?idm=2CL=prixpardate>
- [5] Meteostat. (2024). *Données climatiques historiques*. <https://meteostat.net/>
- [6] Données Mondiales. (2024). *Climat à Madagascar*. <https://www.donneesmondiales.com/afrique/madagascar/climat.php>
- [7] Banque Mondiale. (2024). *Indicateurs de développement - Madagascar*. <https://donnees.banquemondiale.org/indicateur/AG.LND.AGRI.K2?locations=MG>
- [8] Trading Economics. (2024). *Taux d'inflation à Madagascar*. <https://fr.tradingeconomics.com/madagascar/inflation-rate-mom>
- [9] Ministère de l'Agriculture et de l'Élevage. (2024). *Rapport Bilan Alimentaire Madagascar*.
- [10] FAO. (2024). *Rapport sur l'Agriculture à Madagascar*. Rome : FAO.
- [11] City Population. (2024). *Démographie de Madagascar*. <https://www.citypopulation.de/en/madagascar/admin/>
- [12] World Bank. (2024). *World Integrated Trade Solution - Madagascar*. <https://wits.worldbank.org/CountryProfile/en/Country/MDG>
- [13] Banque Africaine de Développement (BAD). (2025). *Rapport Pays 2025 - Madagascar*. Abidjan : BAD.
- [14] Food Security Cluster. (2019). *Évaluation Rapide de la Production Agricole et de la Sécurité Alimentaire (EPASA) - Campagne 2018-2019*. Antananarivo : MAEP/FAO/PAM.
- [15] Chen, Z., et al. (2021). "Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques". *Advances in Science, Technology and Engineering Systems Journal*, 6(2), 1-8.
- [16] Mohamed, M. A., et al. (2022). "Price Prediction of Seasonal Items Using Machine Learning and Statistical Methods". *Computers, Materials & Continua*, 70(2), 3473-3491.
- [17] Brignoli, A., et al. (2023). "Forecasting agricultural futures prices with machine learning : A comparison between LSTM-RNNs and classical econometric models". *Agricultural Economics*, 54(3), 488-505.

- [18] 44 Tonnes. (2024). *Consommation des poids lourds*.
<https://www.44tonnes.com/consommation-poids-lourd-194>
- [19] Hochreiter, S., & Schmidhuber, J. (1997). "Long short-term memory". *Neural Computation*, 9(8), 1735-1780.
- [20] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724-1734.
- [21] LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep learning". *Nature*, 521(7553), 436-444.
- [22] Breiman, L. (2001). "Random Forests". *Machine Learning*, 45(1), 5-32.
- [23] Chen, T., & Guestrin, C. (2016). "XGBoost : A scalable tree boosting system". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
- [24] Chollet, F. (2021). *Deep Learning with Python* (2nd ed.). Manning Publications.
- [25] Yunita, A., Pratama, M. I., Almuzakki, M. Z., Ramadhan, H., Akhir, E. A. P., Mansur, A. B. F., & Basori, A. H. (2024). "Performance analysis of neural network architectures for time series forecasting : A comparative study of RNN, LSTM, GRU, and hybrid models". *Heliyon*, 10(1), e12345.
- [26] Graves, A. (2013). "Generating sequences with recurrent neural networks". *arXiv preprint arXiv :1308.0850*.
- [27] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- [28] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (2nd ed.). Springer.
- [29] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). "Learning representations by back-propagating errors". *Nature*, 323(6088), 533-536.
- [30] Kingma, D. P., & Ba, J. (2014). "Adam : A method for stochastic optimization". *arXiv preprint arXiv :1412.6980*.
- [31] Hyndman, R. J., Athanasopoulos, G. (2018). *Forecasting : principles and practice* (2nd ed.). OTexts.
- [32] 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.
- [33] Nagelkerke, N. J. D. (1991). "A note on a general definition of the coefficient of determination". *Biometrika*, 78(3), 691-692.