Cahier des charges

Cahier des charges de la détection et Caractérisation Automatisées d'Exoplanètes par Intelligence Artificielle

S. Gallais, M.Rolland, C. De Blauwe, M. Leitao, O. Schwartz, K. Benjelloum

I. Contexte & Opportunité

1.1. Contexte

L'astronomie vit une révolution des données grâce aux missions Kepler et TESS de la NASA, qui ont produit des téraoctets de mesures photométriques. Ces courbes de lumière contiennent potentiellement des milliers d'exoplanètes non détectées. Cependant, l'analyse manuelle est devenue impossible face au volume, et les méthodes de Deep Learning (CNN) sont souvent trop lourdes et opaques pour une vérification rapide.

1.2. Opportunité

Il existe un besoin pour un outil intermédiaire : plus puissant que l'analyse manuelle, mais plus léger, rapide et explicable que les réseaux de neurones profonds. L'opportunité est de concevoir un système automatisé basé sur le Machine Learning classique, capable de tourner sur des machines standard, pour démocratiser la recherche d'exoplanètes.

II. Problématique et Besoin

2.1. Problématique

Comment développer un système robuste, automatisé et reproductible capable de détecter, caractériser et visualiser des exoplanètes à partir des courbes de lumière stellaire provenant des missions TESS et Kepler ?

2.2. Besoin

Le système doit permettre de :

1. Récupérer et nettoyer les données brutes de la NASA.
2. Classer les étoiles (Planète vs Pas de Planète) avec un taux de fiabilité élevé.
3. Visualiser les résultats pour permettre une validation humaine.

III. Études Préalables

3.1. Faisabilité Technique

Pour les algorithmes nous nous appuieront sur l'état de l'art (Malik 2022, Ay 2024)qui confirme que des modèles comme XGBoost couplés à une extraction de caractéristiques (*Feature Engineering*) atteignent des performances similaires au Deep Learning (Précision ~98%) pour un coût de calcul minime.

Les données sont publiques (MAST Archive). Le déséquilibre des classes (peu de planètes) sera géré par la génération de données synthétiques (Cuéllar 2022).

Le projet est faisable sur des ordinateurs standards (8-16 Go RAM, CPU classique) grâce au choix du Machine Learning classique.

3.2. Analyse SWOT

Forces	**Faiblesses**
Données NASA de haute qualité.	Les données brutes sont très bruitées.
Modèle XGBoost rapide et interprétable.	Risque de Faux Positifs
Opportunités	**Menaces**
Création d'un outil pédagogique Open Source.	Changements dans l'API MAST.
Validation de nouvelles exoplanètes potentielles.	Difficulté à détecter les très petites planètes (Terres).

IV. Analyse Fonctionnelle

4.1. Fonction Principale

FP1 : Identifier et caractériser des candidats exoplanètes à partir de courbes de lumière stellaire brutes.

4.2. Fonctions secondaires

FS1 - Acquisition : Télécharger les séries temporelles via l'API Lightkurve (Kepler/TESS).

FS2 - Prétraitement : Nettoyer le signal (retrait des outliers, lissage, mise à plat/flattening) et replier la courbe sur sa période (Folding).

FS3 - Enrichissement (Data Augmentation) : Générer des courbes de lumière synthétiques (simulations physiques) pour entraîner le modèle sur des cas rares.

FS4 - Analyse (Machine Learning) :

- Extraire les caractéristiques mathématiques (features) via TSFRESH.
- Classer les courbes via un modèle XGBoost entraîné.

FS5 - Visualisation : Afficher les courbes (brutes/repliées) et le verdict de l'IA sur une interface Web.

V. Contraintes

5.1. Contraintes Techniques

Langages : Python 3.9+ (Backend), React.js (Frontend).

Bibliothèques ML : scikit-learn, xgboost, tsfresh, imbalanced-learn (pour SMOTE).

Astronomie : Lightkurve, Astropy.

Performance : Le traitement d'une étoile doit prendre moins de 10 secondes sur un CPU standard.

5.2. Contraintes Scientifiques

Interprétabilité : Le modèle doit fournir les "features importance" (ex: la profondeur du transit a joué à 80% dans la décision).

Rigueur : Validation croisée obligatoire avec les catalogues d'exoplanètes confirmées (NASA Exoplanet Archive).

5.3. Contraintes Économiques et Légales

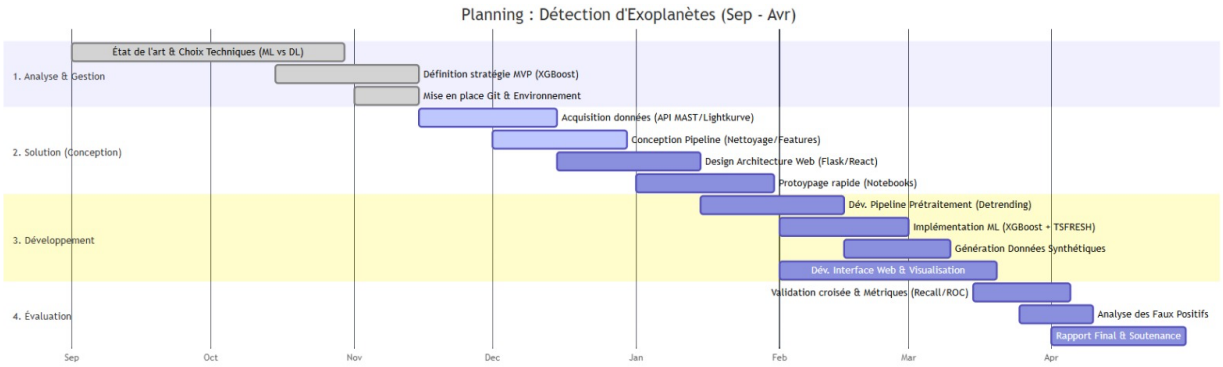
Coût : 0€ (Utilisation exclusive de logiciels Open Source).

Licence : Citation obligatoire des sources NASA/MAST

VI. Exigences fonctionnelles et critères de validation

Exigence	Critère de Validation (Métriques)
ET1. Qualité de la Détection	≥ 90 % de précision sur dataset connu
ET2. Téléchargement des données	Fonctionne sur au moins 100 étoiles
ET3. Prétraitement du signal	Bruit réduit et données normalisées
ET4. Interface web	Visualisation d'au moins 3 étoiles en temps réel
ET5. Sécurité	Authentification fonctionnelle + accès restreint
ET6. Documentation	Tutoriel complet + architecture

VII. Planning prévisionnel



VIII. Annexes

8.1. Matrice RACI

Tâches	R	A	C	I
Pipeline de traitement	Mathis	Oscar	Simon	Tous
Modèle ML	Simon	Oscar	Mathis	Tous
Interface web	Charles	Oscar	Méderic	Tous
Sécurité	Kamil	Oscar	Charles	Tous
Architecture code	Méderic	Oscar	Simon	Tous
Documentation	Oscar	/	Tous	/