État de l’art

État de l’art 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. Introduction

1.1. Contexte Astrophysique et Big Data

La découverte de la première exoplanète en 1995 a ouvert une nouvelle ère en astronomie. Trente ans plus tard, avec plus de 5000 exoplanètes confirmées, la discipline a changé d'échelle. Les missions spatiales de la NASA, *Kepler* (lancée en 2009) et *TESS* (Transiting Exoplanet Survey Satellite, lancé en 2018), ont transformé l'exoplanétologie en une science de données massives ("Big Data"). Kepler a surveillé environ 200 000 étoiles en continu pendant quatre ans, générant des séries temporelles de photométrie d'une précision inégalée. TESS, couvrant 85% du ciel, produit environ un million de courbes de lumière par mois.

1.2. La Problématique Technique : Le Signal dans le Bruit

La méthode des transits, qui consiste à détecter la baisse périodique de luminosité d'une étoile lorsqu'une planète passe devant elle, est la plus prolifique. Cependant, elle souffre d'un taux élevé de faux positifs. Les variations de luminosité peuvent être causées par des étoiles binaires à éclipses, la variabilité stellaire naturelle, ou du bruit instrumental.Historiquement, l'analyse reposait sur des algorithmes comme le *Box Least Squares* (BLS) suivis d'une inspection humaine manuelle ("vetting"). Face au volume de données actuel, cette approche n'est plus viable. L'enjeu est de développer des systèmes automatisés capables non seulement de détecter des signaux périodiques (Threshold Crossing Events - TCEs), mais surtout de classer ces signaux avec une fiabilité comparable ou supérieure à celle des experts humains.

1.3. Périmètre de l'Étude

Cet état de l'art analyse les méthodes d'apprentissage automatique (*Machine Learning* et *Deep Learning*) appliquées à cette problématique entre 2018 et 2025. Nous nous concentrons sur l'analyse des courbes de lumière (1D), tout en examinant les avancées récentes en imagerie directe (3D) pour leurs apports en traitement du signal.

II. Développement Thématique : Évolution des Architectures

2.1. L'Ère du Deep Learning : Réseaux Convolutionnels et Représentations

L'année 2018 a marqué un tournant avec les travaux de Shallue & Vanderburg, qui ont établi les Réseaux de Neurones Convolutionnels (CNN) comme le standard industriel pour le traitement des données Kepler. Leur approche se distingue par un prétraitement spécifique des courbes de lumière : au lieu d'injecter la série temporelle brute, ils génèrent deux "vues" distinctes. La première, une vue globale de la courbe repliée sur la période orbitale, permet au réseau d'appréhender la forme générale du signal et de repérer d'éventuelles éclipses secondaires. La seconde, une vue locale centrée sur le transit, offre une résolution suffisante pour analyser la géométrie précise de la baisse de luminosité. Cette architecture a permis d'atteindre une précision de 96% sur le jeu de test Kepler.

Cependant, cette méthode se heurte à une limite structurelle : la rareté des exoplanètes confirmées, qui restreint la base d'entraînement des réseaux de neurones. Pour contourner cet obstacle, Cuéllar et al. (2022) ont proposé une stratégie d'augmentation de données. Leur méthode repose sur l'injection massive de signaux synthétiques, générés à partir de modèles physiques (comme celui de Mandel & Agol), dans des courbes de lumière réelles. Ils démontrent qu'un modèle entraîné avec une majorité de données synthétiques (environ 74%) généralise mieux sur les données réelles. De plus, ils innovent en transformant les courbes 1D en images 2D, empilant les périodes les unes sur les autres, ce qui aide le réseau à distinguer la cohérence temporelle d'un transit face au bruit aléatoire.

2.2. La résurgence du Machine Learning classique : efficacité et explicabilité

En réaction à la lourdeur des modèles de Deep Learning, qui nécessitent souvent des infrastructures de calcul coûteuses (GPU) et fonctionnent comme des "boîtes noires", une seconde école de pensée prône un retour à des algorithmes plus légers. Les travaux de Malik et al. (2022) illustrent parfaitement cette tendance. Plutôt que de laisser un réseau de neurones extraire lui-même les caractéristiques du signal, ils utilisent des bibliothèques d'ingénierie des fonctionnalités, telles que *TSFRESH*, pour calculer mathématiquement près de 800 descripteurs par courbe (entropie, kurtosis, coefficients de Fourier). Ces données alimentent ensuite des modèles d'arbres de décision boostés (type *LightGBM*).

Cette approche présente des avantages opérationnels majeurs. D'une part, le temps d'entraînement passe de plusieurs heures à quelques minutes sur un processeur standard. D'autre part, les performances restent compétitives, avec une aire sous la courbe ROC (AUC) de 0.948 sur Kepler, très proche des standards du Deep Learning. Ay (2024) confirme la pertinence de cette stratégie, notamment pour des jeux de données de taille restreinte. Ses comparaisons montrent que sur des échantillons limités (~5000 étoiles), les algorithmes comme *XGBoost* ou *AdaBoost* surpassent les réseaux de neurones profonds, ces derniers ayant tendance à faire du sur-apprentissage sur le bruit.

2.3. Vers des approches hybrides : quand l'IA rencontre la physique

Bien que ce projet se concentre sur la méthode des transits, les avancées récentes en imagerie directe offrent des perspectives cruciales sur le traitement du bruit non-stationnaire. Les recherches de Flasseur et al. (2024) mettent en évidence les limites du Deep Learning lorsqu'il est appliqué directement sur des données brutes très bruitées. Ils introduisent le concept de *Deep PACO*, une méthode hybride où les données subissent d'abord un "blanchiment" statistique rigoureux pour supprimer les corrélations spatiales du bruit avant d'être traitées par un réseau U-Net.

Plus récemment, Bodrito et al. (2025) ont poussé cette logique plus loin avec l'architecture *ExoMILD*. Au lieu de traiter l'image comme une simple matrice de pixels, leur réseau intègre explicitement des connaissances physiques, telles que les symétries de rotation et les échelles spatiales propres à l'instrument. Cette fusion entre modélisation physique et apprentissage profond permet de combiner plusieurs nuits d'observation pour détecter des orbites complètes, augmentant drastiquement la sensibilité de détection. Ces travaux suggèrent que pour les transits les plus faibles, l'avenir réside probablement dans des architectures guidées par la physique plutôt que dans le "tout neuronal".

III. Discussion

L'analyse croisée de la littérature met en lumière une tension entre la complexité des modèles et la qualité des données disponibles.

Si les Réseaux de Neurones (CNN) restent la référence pour leur capacité à capturer des motifs subtils dans de grands volumes de données, ils souffrent d'un manque d'interpretabilité et d'une forte dépendance à la quantité de données annotées. À l'inverse, les méthodes d'ensemble (XGBoost, LightGBM) offrent un excellent compromis entre rapidité, robustesse et explicabilité, ce qui en fait des candidats idéaux pour un prototypage rapide ou des ressources de calcul limitées.

Enfin, un consensus émerge sur l'importance critique du prétraitement. Que ce soit par la génération de données synthétiques pour équilibrer les classes (Cuéllar) ou par le blanchiment statistique du bruit (Flasseur), la qualité de la donnée d'entrée s'avère souvent plus déterminante que le choix de l'architecture du réseau lui-même.

3.1. Matrice de Comparaison Technique

Critère	**Deep Learning (CNN 1D/2D)**	**Machine Learning Classique (XGBoost)**	**Approche Hybride (Stats + DL)**
Références Clés	Shallue (2018), Cuéllar (2022)	Malik (2022), Ay (2024)	Flasseur (2024), Bodrito (2025)
Entrée du Modèle	Données brutes normalisées (Vecteurs/Images)	Vecteur de caractéristiques extraites	Données "blanchies" statistiquement
Performance (Grand Dataset)	Excellente	Très bonne, mais peut saturer	Potentiellement supérieure
Performance (Petit Dataset)	Faible	Excellente	Bonne
Coût de Calcul	Élevé	Faible	Élevé

IV. Conclusion et Orientations pour le Projet

Au terme de cette analyse bibliographique, l'adoption d'une approche fondée exclusivement sur le Machine Learning classique s'impose comme la stratégie la plus pertinente pour notre projet. L'état de l'art démontre en effet que sur des tâches de classification binaire de transits, la plus-value du Deep Learning reste marginale face à des modèles d'ensemble bien calibrés, alors que sa complexité de mise en œuvre et son coût computationnel sont nettement supérieurs.

Par conséquent, nous concentrerons nos efforts sur le développement d'un pipeline robuste associant une ingénierie des fonctionnalités rigoureuse à un algorithme de type XGBoost. Ce choix technologique

nous permet de garantir une interprétabilité physique des résultats, indispensable pour la validation scientifique, tout en conservant une agilité de développement compatible avec les délais du projet.

La performance de ce modèle ne reposera pas sur la complexité de son architecture, mais sur la qualité des données : nous intégrerons impérativement l'enrichissement du jeu d'entraînement par des données synthétiques, seule méthode viable pour corriger le déséquilibre structurel des catalogues astronomiques sans recourir à l'opacité des réseaux de neurones profonds.

V. Références Bibliographiques

[1]. Shallue, C. J., & Vanderburg, A. (2018). Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90.

[2]. Cuéllar, S., et al. (2022). Deep learning exoplanets detection by combining real and synthetic data.

[3]. Malik, A., Moster, B. P., & Obermeier, C. (2022). Exoplanet detection using machine learning. *Monthly Notices of the Royal Astronomical Society*.

[4]. Ay, B. (2024). Comparative Analysis of Machine Learning and Neural Network Approaches for Exoplanet Identification. *MSc Research Project*, National College of Ireland.

[5]. Flasseur, O., et al. (2024). Deep PACO: combining statistical models with deep learning for exoplanet detection and characterization in direct imaging at high contrast. *Monthly Notices of the Royal Astronomical Society*,

[6]. Bodrito, T., et al. (2025). ExoMILD: Deep Learning for Exoplanet Detection and Characterization by Direct Imaging at High Contrast.