

Défi IA 2025 :
**Trouver de l'eau et des nuages
dans les atmosphères d'exoplanètes**

Auteur
DRISS CHRAIBI



Table des matières

1	Introduction et Contexte	2
1.1	Enjeux de la Mission ARIEL	2
1.2	Objectifs du Défi	2
2	Analyse et Prétraitement des Données	2
2.1	Description du Jeu de Données	2
2.2	Stratégie d'Augmentation de Données (Data Augmentation)	2
2.3	Domain Shift et difficultés rencontrés lors de la normalisation	2
2.4	Ingénierie des Caractéristiques (Feature Engineering)	3
3	Méthodologie et Modèles	4
3.1	Algorithmes Utilisés	4
3.2	Protocole d'Entraînement	4
3.3	Prédiction et Stratégie d'Ensemble	5
4	Explicabilité et Analyse du Modèle	5
4.1	Dualité des Stratégies de Détection	6
4.2	Contextualisation et Interactions Non-Linéaires	6
4.3	Choix de la Parsimonie	6
4.4	Limites Intrinsèques des Données	7
5	Conclusion	7

1 Introduction et Contexte

1.1 Enjeux de la Mission ARIEL

La mission ARIEL (Atmospheric Remote-sensing Infrared Exoplanet Laboratory) est une mission spatiale de l'ESA destinée à étudier les atmosphères des exoplanètes. Prévue pour 2029, ARIEL vise à révéler les compositions chimiques et physiques de ces atmosphères ainsi que leur structure, en observant des planètes autour d'étoiles différentes du Soleil. Ces observations permettront de mieux comprendre la diversité des exoplanètes et les conditions nécessaires pour la formation et l'évolution des mondes habitables.

1.2 Objectifs du Défi

Les données de ce Défi IA sont des spectres simulés d'ARIEL représentant le produit final des observations de transits planétaires. L'objectif est de trouver dans ces spectres la présence ou l'absence d'eau et de nuage. C'est donc deux classification binaires.

2 Analyse et Prétraitement des Données

2.1 Description du Jeu de Données

Les données d'entrée pour ce Défi IA 2025 consistent en des spectres infrarouges et visibles de 52 points, capturés par les instruments de la mission ARIEL. Chaque spectre est accompagné de cinq points de données auxiliaires décrivant les caractéristiques du système planétaire, telles que la distance à l'étoile hôte, la température de l'étoile, la masse de la planète, et d'autres paramètres astrophysiques pertinents. Ces données auxiliaires fournissent un contexte essentiel pour l'interprétation des spectres, permettant aux modèles d'IA de mieux comprendre les conditions environnementales dans lesquelles les atmosphères exoplanétaires évoluent. L'objectif est de transformer ces spectres et ces données auxiliaires en informations exploitables pour prédire la présence ou l'absence de deux composants clés : l'eau et les nuages.

2.2 Stratégie d'Augmentation de Données (Data Augmentation)

Pour limiter le sur-apprentissage sur le bruit instrumental, nous avons appliqué une augmentation dynamique durant l'entraînement :

$$S'_i = S_i + \mathcal{N}(0, \alpha \cdot \sigma_i) \quad (1)$$

où S_i est le signal spectral, σ_i l'incertitude instrumentale, et α un facteur d'échelle (fixé à 0.5). Les données auxiliaires restent inchangées.

Cela nous permet d'augmenter les données autant que l'on veut, de manière réaliste puisqu'on utilise directement l'incertitude instrumentale pour déterminer la variance du bruit synthétique.

2.3 Domain Shift et difficultés rencontrés lors de la normalisation

Il est nécessaire de normaliser les données, puisque les features ont des ordres de grandeurs très différents. Le choix de la normalisation a été d'une grande importance. En effet, la "baseline" nous invitait à une normalisation par max, qui permet d'avoir toutes nos données dans l'intervalle [0,1]. Cependant, une telle normalisation est très sensible aux outliers. Elle tend également à supprimer l'information liée à l'intensité lumineuse absolue afin de ne conserver que la forme relative du signal spectral. Or, cette intensité constitue une information discriminante importante pour la classification des nuages, qui agissent comme des couches absorbantes et diffusantes qui atténuent le flux observé.

Idée 1 : Normalisations adaptée à chaque classe

En appliquant une normalisation StandardScaler spécifiquement à la classe nuage, les performances en train/validation ont progresser immédiatement. Cependant, Les distributions train/validation et test montre un domain shift qui empêche une généralisation correct. Il est donc nécessaire d'avoir une normalisation indépendante des données d'entraînement

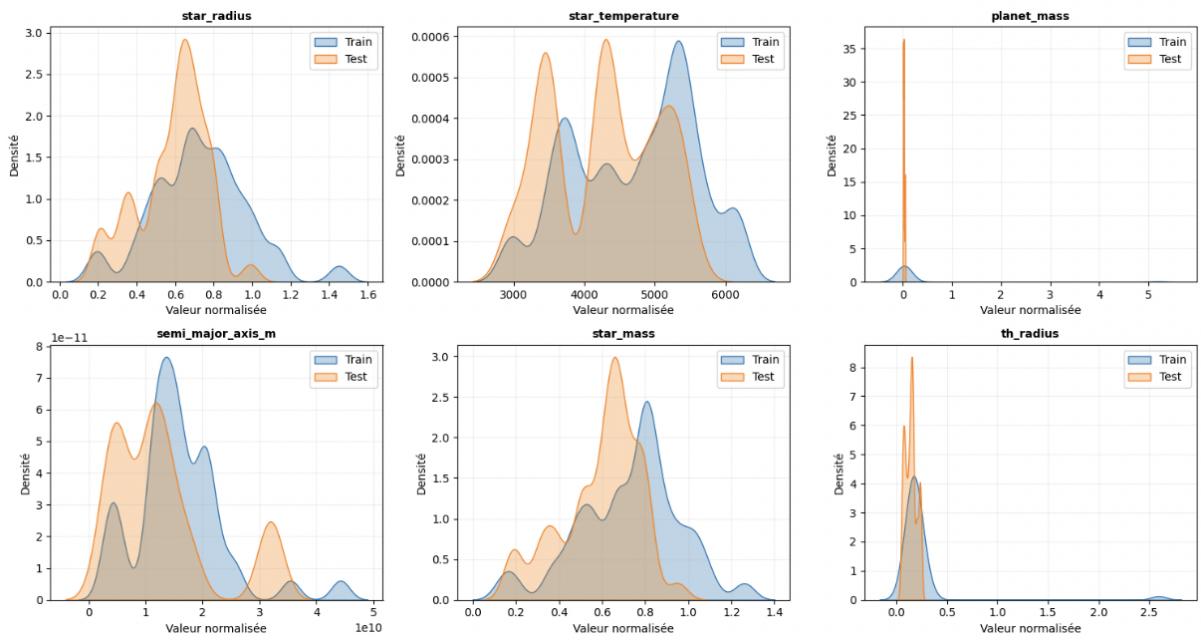


Fig. 1. Domain shift entre train et test

Idée 2 : Remplacement du max par la norme L2 et ajouts de features

Pour pallier le problème du Domain Shift tout en capturant les nuances des nuages, nous avons mis en place une stratégie hybride : Normalisation L2 (Sample-wise) + Ajout de nouvelles features pour ne pas perdre d'informations.

Nous avons remplacé la normalisation Max par une normalisation vectorielle L2. Chaque spectre x est divisé par sa norme euclidienne $\|x\|_2$.

$$x_{norm} = \frac{x}{\sqrt{\sum x_i^2}}$$

Cette approche présente deux avantages : elle est calculée individuellement pour chaque spectre (donc insensible au Domain Shift global) et elle est beaucoup plus robuste aux pics de bruit isolés que la normalisation par le Max.

2.4 Ingénierie des Caractéristiques (Feature Engineering)

Puisque la normalisation L2 projette les vecteurs sur une hypersphère unitaire (effaçant l'information de luminosité absolue), nous avons calculé des statistiques descriptives sur le signal brut avant normalisation et les avons ajoutées comme variables auxiliaires. Nous avons ainsi injecté :

- La Moyenne et l'Écart-type (pour capturer l'atténuation globale et le contraste).
- Le Max et le Min (pour l'amplitude).
- Des ratios spectraux pour caractériser la profondeur des raies d'absorption indépendamment de l'échelle.

Cette combinaison permet au modèle XGBoost de bénéficier du meilleur des deux mondes : la forme propre du spectre (via L2) pour détecter les signatures chimiques (H_2O), et l'énergie brute (via les features ajoutées) pour identifier l'opacité des nuages.

3 Méthodologie et Modèles

Notre approche repose sur une stratégie d'ensemble (*Ensemble Learning*) combinant des architectures complémentaires.

3.1 Algorithmes Utilisés

Dans le cadre de notre approche d'apprentissage d'ensemble (*Ensemble Learning*), nous avons combiné deux familles d'algorithme aux philosophies complémentaires :

- **RandomForest** : Basé sur le principe du *Bagging*, cet algorithme construit une forêt d'arbres de décision indépendants et entraînés en parallèle. En moyennant les prédictions de multiples arbres décorrélés, il réduit la variance du modèle. Il assure une grande robustesse au bruit instrumental et permet de généraliser efficacement sur les structures globales (comme la texture des nuages), sans se focaliser sur les outliers. Il a été particulièrement adapté étant donné le domain shift constaté précédemment entre train/validation et test.
- **XGBoost** : Basé sur le principe du *Gradient Boosting*, cet algorithme construit des arbres de manière séquentielle, où chaque nouvel arbre cherche à corriger les erreurs résiduelles des précédents. *Rôle* : Grâce à sa capacité à optimiser une fonction de perte spécifique, il réduit le biais du modèle. Il est crucial pour capturer les fines structures non-linéaires, telles que les faibles raies d'absorption de l'eau (H_2O) que le bruit ambiant tend à masquer.
- **ResNet-CBAM Hybride (Architecture Profonde)** : Contrairement aux approches classiques, nous avons développé une architecture neuronale hybride à double entrée.

Architecture : Le modèle combine une branche convolutionnelle profonde (ResNet 1D) pour l'analyse spectrale et une branche dense pour les données physiques scalaires.

Innovation (Attention CBAM) : Chaque bloc résiduel intègre un module d'attention *CBAM* (*Convolutional Block Attention Module*). Ce mécanisme permet au réseau de se focaliser dynamiquement sur les canaux pertinents (Channel Attention) et les zones précises du spectre (Spatial Attention), ignorant ainsi le bruit de fond pour cibler les faibles raies d'absorption (ex: H_2O).

Optimisation : Pour maximiser la convergence, nous utilisons l'activation *Swish* (plus performante que ReLU sur les réseaux profonds), l'optimiseur *AdamW* pour une meilleure généralisation, et une stratégie de *Pooling Hybride* (Moyenne + Max) pour ne perdre aucune information sur les pics d'intensité.

3.2 Protocole d'Entraînement

Afin de garantir la robustesse de nos résultats et d'éviter le sur-apprentissage (*overfitting*) sur notre jeu de données limité (3000 spectres), nous avons mis en place une stratégie de validation rigoureuse. La performance de notre modèle XGBoost (98% de précision) a été validée via une procédure de **Validation Croisée Stratifiée à 5 plis** (*Stratified 5-Fold Cross-Validation*).

Ce protocole se déroule selon les étapes suivantes :

- **Segmentation Stratifiée ($K = 5$)** : L'ensemble des données d'entraînement est divisé en 5 sous-ensembles distincts. L'aspect *stratifié* est crucial ici : il garantit que la proportion de chaque classe cible (H_2O et Nuages) est rigoureusement identique dans chaque pli, préservant ainsi la distribution physique originale des données.

- **Cycle d'Apprentissage** : Le processus est répété 5 fois. À chaque itération :
 - Le modèle est entraîné sur 4 plis (soit 80% des données).
 - Le modèle est évalué sur le 5^{ème} pli restant (20% des données), qui agit comme un jeu de test inédit ("Hold-out set").
- **Mécanisme d'Arrêt Précoce (*Early Stopping*)** : Pour chaque pli, nous surveillons la performance du XGBoost sur le jeu de validation en temps réel. Si la fonction de perte (*LogLoss*) ne s'améliore pas pendant 300 itérations consécutives, l'entraînement est stoppé et nous conservons les poids du modèle ayant obtenu le meilleur score. Cela empêche le modèle de mémoriser le bruit instrumental.

3.3 Prédiction et Stratégie d'Ensemble

Le passage de RandomForest à Gradient Boosting (XGBoost) nous a permis de franchir un premier cap significatif, faisant passer la précision de 95% à 97%. Contrairement au RandomForest qui moyenne les décisions, le Gradient Boosting se focalise spécifiquement sur les échantillons difficiles à classer, permettant d'affiner la frontière de décision là où le signal est le plus complexe. Cette amélioration est particulièrement notable sur la classe "nuage", source principale d'erreurs initialement. Après une optimisation fine des hyperparamètres, le modèle XGBoost seul a atteint un score de **98.2%**.

Cependant, pour maximiser la robustesse du système, nous avons mis en œuvre une stratégie de *Blending* (mélange pondéré) combinant notre meilleur modèle arborescent (XGBoost) et notre architecture neuronale profonde (ResNet-CBAM).

- **La méthode** : Nous avons calculé une moyenne pondérée des probabilités prédites par les deux modèles selon la formule :

$$P_{final} = 0.5 \times P_{XGBoost} + 0.5 \times P_{ResNet}$$

- **Résultat** : Cette approche nous a permis d'atteindre une précision finale de **98.5%**.

Analyse des performances et perspectives

Il est important de noter que cette stratégie de *Blending* n'a pas été pleinement optimisée durant la compétition. En effet, nos expérimentations initiales affichaient un score modéré de **97.8%** sur le classement public (*Public Leaderboard*), ce qui nous a dissuadés de complexifier davantage le mélange.

La performance réelle a "explosé" sur le classement privé (*Private Leaderboard*), atteignant **98.5%**. Cet écart significatif suggère que notre modèle généralise bien mieux que prévu et que nous avons sous-estimé son potentiel. Avec le recul, nous aurions pu améliorer ce résultat :

- En optimisant conjointement les hyperparamètres des deux architectures (XGBoost et ResNet) spécifiquement pour le mélange.
- En remplaçant la pondération manuelle (0.5/0.5) par un *méta-modèle* (technique de *Stacking*), comme une régression logistique, qui aurait appris automatiquement les poids optimaux pour chaque classe, tirant ainsi le meilleur parti de la complémentarité des deux modèles.

Ce gain final démontre la *complémentarité* des deux approches : XGBoost excelle sur les données structurées et les frontières de décision nettes, tandis que le ResNet capture des subtilités spectrales via ses mécanismes d'attention, corrigeant certaines erreurs résiduelles du Boosting.

4 Explicabilité et Analyse du Modèle

Au-delà de la performance métrique brute, nous avons cherché à valider la cohérence structurelle du modèle. L'analyse des valeurs SHAP (SHapley Additive exPlanations), illustrée ci-dessous, permet de

vérifier que le modèle ne repose pas sur des artefacts d'apprentissage, mais exploite des signaux structurels distincts pour chaque classe.

4.1 Dualité des Stratégies de Détection

L'analyse comparée des Figures 2a et 2b révèle que le modèle adapte sa stratégie d'extraction de caractéristiques selon la nature physique de la cible :

- **Pour H₂O (Triangulation Spectrale) :** L'analyse des features dominantes (Figure 2b) montre que le modèle focalise son attention sur trois zones clés : le cœur de l'absorption (Spec_50) et deux zones de référence distantes (Spec_26 et Spec_46). Cette sélection n'est pas aléatoire. Le modèle utilise Spec_26 et Spec_46 comme points d'ancrage du "continuum" (zones où l'absorption est moindre) pour mesurer la profondeur relative du creux en Spec_50. Le modèle utilise une détection de motif géométrique.
- **Pour les Nuages (Analyse de Distribution) :** À l'inverse, la Figure 2a montre une importance accrue des statistiques descriptives globales (Mean, Skewness). Contrairement au gaz qui crée un motif local, les nuages altèrent la distribution globale du signal. Le modèle a appris à corrélérer un signal "plat" (faible variance) et atténué (faible moyenne) à la présence d'obstruction atmosphérique.

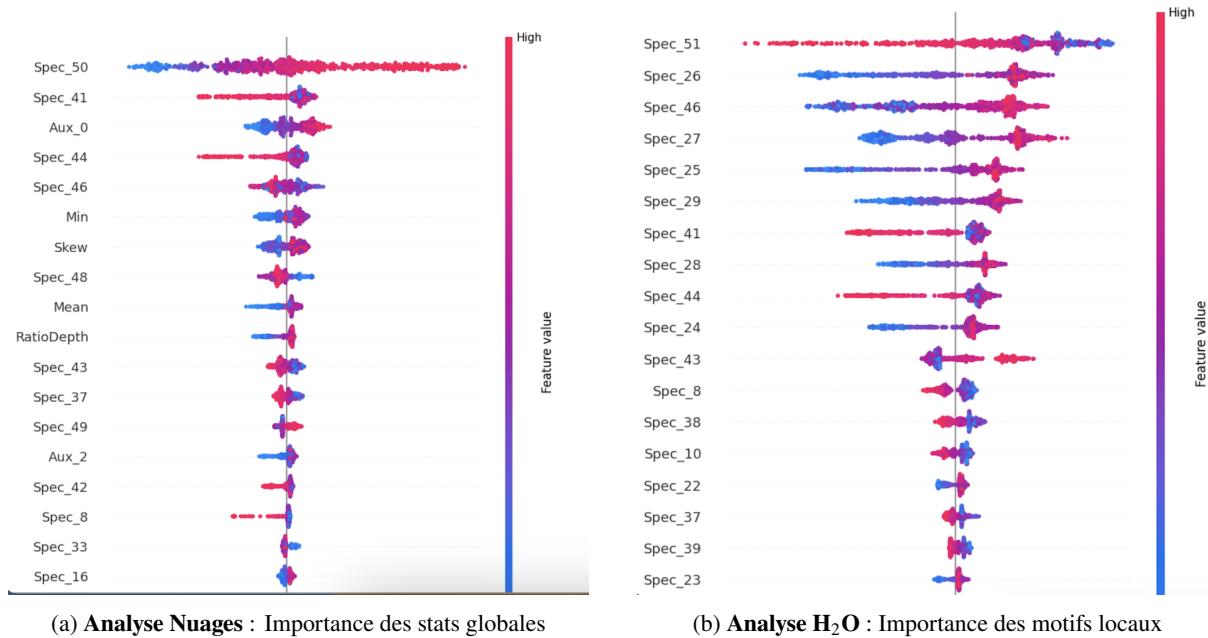


Fig. 2. Comparaison des distributions SHAP pour les deux cibles.

4.2 Contextualisation et Interactions Non-Linéaires

Un point notable de l'analyse est l'importance significative des métadonnées exogènes (star_radius, planet_mass) pour la prédiction des nuages. Cela indique que le modèle ne traite pas le spectre de manière isolée mais contextuelle. Il a appris des interactions non-linéaires complexes : la "signature" d'un nuage (spectre plat) n'a pas la même signification statistique selon la taille de l'étoile ou la température du système. Le modèle ajuste donc implicitement ses seuils de décision en fonction des propriétés physiques de l'objet observé.

4.3 Choix de la Parsimonie

Lors de la phase d'optimisation, nous avons confronté une approche complexe (incluant dérivées secondees et Z-scores) à une approche robuste (spectres bruts et statistiques simples). Bien que l'approche complexe

ait montré des gains sur le jeu d'entraînement, elle a stagné sur le jeu de test. Nous avons attribué ce phénomène à un sur-apprentissage du bruit et à une sensibilité au *Covariate Shift* (décalage de distribution des métadonnées entre Train et Test).

4.4 Limites Intrinsèques des Données

Enfin, l'analyse des erreurs résiduelles montre une forte corrélation avec le Rapport Signal-sur-Bruit (SNR). Les échecs se concentrent systématiquement dans deux zones : les signaux de très faible intensité (indiscernables du bruit de fond) et les spectres très intenses (sujets à une forte variance statistique). Ces erreurs reflètent les limites physiques de l'instrumentation plutôt qu'une défaillance algorithmique.

5 Conclusion

Ce travail a illustré l'importance de la connaissance physique pour l'enrichissement des modèles d'IA. L'intuition qui a guidé ce travail a été, dès le début, de privilégier les méthodes d'ensemble qui sont connues pour exceller sur les petits datasets, mais en les enrichissant par la connaissance physique à l'aide de l'ingénierie des caractéristiques pour limiter la perte d'informations liée à la normalisation. C'est véritablement cet ajout qui a permis d'augmenter drastiquement les performances de nos modèles, notamment pour la prédiction des nuages.

La dernière partie illustre un autre grand avantage des modèles d'ensemble : leur interprétabilité par rapport aux modèles de réseaux de neurones profonds. L'analyse d'explicabilité a confirmé que le modèle ne se contentait pas d'apprendre par cœur, mais qu'il reproduisait des raisonnements astrophysiques valides (triangulation spectrale, contextes stellaires), offrant ainsi les garanties de robustesse nécessaires à l'analyse scientifique.

Cependant, au lieu d'opposer les méthodes d'ensemble aux méthodes "Deep Learning", ce travail a montré qu'allier les deux permettait également de construire des modèles plus robustes et plus généralisables. Pour aller plus loin, une piste d'amélioration naturelle consisterait à remplacer notre pondération manuelle par une véritable stratégie de *Stacking*. L'utilisation d'un méta-modèle pour apprendre la combinaison optimale des prédictions permettrait sans doute de tirer encore mieux parti des spécificités de chaque architecture, ouvrant la voie vers des systèmes de classification encore plus autonomes et performants.

References

- [1] Tinetti, G., et al. (2018). *A chemical survey of exoplanets with ARIEL*. Experimental Astronomy, 46, 135-209.
- [2] Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
- [3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778).