# Báo cáo Ariel Data Challenge 2025: Xây dựng mô hình học sâu dự đoán phổ quá cảnh

Đỗ Đức Thắng
Khoa Công nghệ Thông tin, ĐHQGHN - UET
Hà Nội, Việt Nam
23020158@vnu.edu.vn

***Tóm tắt nội dung*—Đặc trưng hóa bầu khí quyển của các ngoại hành tinh từ phổ quá cảnh là một bài toán phức tạp do dữ liệu đo đạc thường chứa nhiễu nhiễu và hiệu ứng thiết bị. Cuộc thi Ariel Data Challenge 2025 do NeurIPS và Cơ quan Vũ trụ châu Âu (ESA) tổ chức cung cấp bộ dữ liệu mô phỏng thực tế nhằm thúc đẩy việc áp dụng các kỹ thuật học máy trong lĩnh vực này. Báo cáo này trình bày mô hình dự đoán phổ quá cảnh dựa trên *Residual Network* (ResNet), sử dụng đầu vào gồm *white light transit depth* và các đặc trưng vật lý của hành tinh – sao chủ. Quy trình gồm tiền xử lý tín hiệu, tính toán độ sâu phổ trắng bằng tối ưu hóa, và tinh chỉnh giá trị tại bước sóng $0.7\,\mu\mathrm{m}$ – vùng có ảnh hưởng lớn tới điểm số đánh giá. Kết quả cho thấy ResNet với 80 khối residual giảm sai số MAE tại $0.7\,\mu\mathrm{m}$ khoảng 40,3% so với phương pháp baseline, đồng thời duy trì độ ổn định cao với độ bất định ước lượng chỉ 0.0009. Mô hình đề xuất minh chứng khả năng kết hợp thông tin vật lý và kiến trúc học sâu để cải thiện độ chính xác dự đoán phổ quá cảnh.**

***Index Terms*—Residual Network, học sâu, ước lượng bất định, PyTorch**

## I. Giới thiệu

Đặc trưng hóa bầu khí quyển các hành tinh ngoài hệ Mặt Trời là một trong những bài toán trọng yếu của thiên văn học hiện đại. Việc trích xuất chính xác các thành phần hóa học từ phổ quá cảnh (transit spectrum) quan sát được đóng vai trò then chốt để hiểu về tiến hóa hành tinh, thành phần khí quyển cũng như khả năng tồn tại sự sống. Cuộc thi Ariel Data Challenge 2025, thuộc chuỗi kì thi được NeurIPS tổ chức để thử nghiệm các phương pháp học máy xử lý tín hiệu phổ mô phỏng với độ nhiễu cao và nhiều hiệu ứng vật lý giống thiết bị thực tế.

Trong cuộc thi, tôi phát triển một mô hình dự đoán phổ dựa trên các trụ cột chính:

- Tiền xử lý dữ liệu (Signal pre-processing): Làm sạch dữ liệu, loại bỏ nhiễu do thiết bị đo đạc và tính toán White light transit depth.
- Cải thiện độ chính xác bằng Residual Network (ResNet): Mô hình sử dụng dữ liệu đầu vào gồm độ sâu phổ trắng và các đặc trưng của hành tinh để dự đoán phổ quá cảnh.
- Ước lượng độ không chắc chắn của dự đoán

## II. Dữ liệu

- Đối với mỗi hành tinh, dữ liệu được cho gồm có:
  - Tín hiệu FGS1:(FGS1_signal_0.parquet) với kích thước [135000, 32 × 32] (135000 bước thời gian, mỗi bước tương ứng với 0,1 giây và 32 × 32 là dữ liệu từ cảm biến)
  - Tín hiệu AIRS-CH0: (AIRS-CH0_signal_0.parquet) với kích thước [11250, 32 × 356] (11250 bước thời gian, mỗi bước tương ứng với 1,2 giây và 32 × 356 là dữ liệu từ cảm biến, gồm 32 chiều không gian và 356 bước sóng khác nhau)
- Bộ dữ liệu hiệu chỉnh gồm các thành phần nhằm hiệu chỉnh số liệu giúp giảm thiểu nhiễu trong quá trình đo.
- Các đặc trưng của hành tinh gồm mã định danh, bán kính ngôi sao, khối lượng ngôi sao, nhiệt độ ngôi sao, khối lượng hành tinh, độ lệch tâm quỹ đạo, chu kỳ quỹ đạo, bán kính trục lớn, độ nghiêng quỹ đạo.

## III. Phương pháp

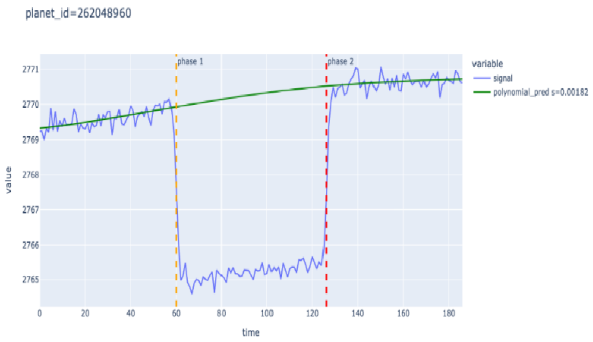
### *3.1 Tiền xử lý dữ liệu*

- Mô hình áp dụng quy trình tiền xử lý dữ liệu theo các bước:
  - ADC (Analog-to-Digital Conversion): chuyển đổi tín hiệu thô sang giá trị số (đơn vị electrons) bằng hệ số gain và offset
  - Loại bỏ dòng tối (Dark current subtraction): xác định và loại bỏ pixel lỗi (hot/dead) để loại bỏ tín hiệu nền không mong muốn từ cảm biến.
  - Hiệu chỉnh phi tuyến (Linearity correction): hiệu chỉnh phi tuyến sử dụng polynomial model từ dữ liệu
  - Loại bỏ dark-current (Dark-Frame Subtraction): Trừ ảnh tối đã được hiệu chỉnh (loại bỏ pixel chết) khỏi ảnh gốc theo thời gian phơi sáng để loại bỏ tín hiệu nền không mong muốn từ cảm biến.
  - Lấy ảnh hiệu chỉnh đôi (Correlated Double Sampling): Tính hiệu giữa ảnh cuối và ảnh đầu mỗi chu kỳ phơi sáng để loại bỏ nhiễu và ổn định tín hiệu.
  - Gộp theo thời gian (Time Binning): Gộp nhiều khung hình theo thời gian để giảm dung lượng và làm mượt chuỗi thời gian quan sát.
  - Hiệu chỉnh trường phẳng (Flat Field Correction) Chia mỗi pixel cho giá trị tương ứng trong ảnh trường phẳng để hiệu chỉnh sự không đồng đều trong độ nhạy của từng pixel

– Loại bỏ các bước sóng ở ngoài rìa và tại các thời điểm đo đầu và cuối để tránh nhiễu, chỉ giữ lại các bước sóng ở trung tâm

### *3.2 Tính white light transit depth*

Phương pháp tính white light transit depth được tham khảo từ notebook Kaggle [1].




Hình 1. Ví dụ về hành tinh id: 28204960

– Phương pháp Tính toán white light transit depth bằng Tối ưu hóa

∗ Nguyên lý cốt lõi là mô hình hóa đường cong ánh sáng quan sát được như một hợp thành của hai thành phần chính:
(1) Sự biến đổi nhỏ của độ sáng ngôi sao theo thời gian
(2) Sự giảm đột ngột độ sáng gây ra bởi sự kiện quá cảnh.

∗ Quy trình bao gồm ba bước chính: xác định các pha quá cảnh, xây dựng hàm mục tiêu dựa trên mô hình khử xu hướng, và tối ưu hóa số để tìm tham số độ sâu.

· Xác định các Pha Quá cảnh.
Để tách riêng tín hiệu quá cảnh, cần xác định hai thời điểm đặc trưng: bắt đầu ($T_1$) và kết thúc ($T_2$) của sự kiện. Các mốc này được xác định bằng cách phân tích đạo hàm bậc nhất của chuỗi thời gian quang thông $F(t)$.
Thời điểm bắt đầu $T_1$ là nơi đạo hàm đạt cực tiểu:

$$\frac{dF}{dt} \to \min$$

Thời điểm kết thúc $T_2$ là nơi đạo hàm đạt cực đại:

$$\frac{dF}{dt} \to \max$$

· Mô hình hóa
Mô hình giả định rằng tín hiệu quang thông trong vùng quá cảnh, ký hiệu là $F_{\text{in-transit}}$, bị suy giảm một cách có hệ thống so với tín hiệu ngoài vùng quá cảnh $F_{\text{out-of-transit}}$. white light transit depth, ký hiệu là $s$, được định nghĩa là hệ số tỷ lệ mô tả mức độ suy giảm này.
Để tìm giá trị $s$, tôi xây dựng một tín hiệu đã hiệu chỉnh $F'(t, s)$, bằng cách áp dụng hệ số nhân $(1 + s)$ cho các điểm dữ liệu nằm trong vùng quá cảnh. Biểu thức được định nghĩa như sau:

$$F'(t,s) = \begin{cases} F(t) \cdot (1+s), & \text{nếu } T_1 \le t \le T_2 \\ F(t), & \text{nếu } t < T_1 \text{ hoặc } t > T_2 \end{cases}$$

Sự biến thiên nội tại của ngôi sao được giả định là một quá trình chậm, có thể mô hình hóa hiệu quả bằng một đa thức bậc thấp $P(t)$. Quá trình khớp và loại bỏ thành phần biến thiên này khỏi tín hiệu được gọi là khử xu hướng . Nếu tham số $s$ được lựa chọn chính xác, tín hiệu đã hiệu chỉnh $F'(t, s)$ sẽ chỉ còn lại nhiễu và thành phần dao động chậm của sao, do đó sẽ khớp tốt với mô hình $P(t)$.

· Tính s bằng tối ưu hóa
Mục tiêu của phương pháp là tìm giá trị tối ưu cho white light transit depth $s_{\text{opt}}$, sao cho tín hiệu sau khi hiệu chỉnh và khử xu hướng khớp tốt nhất với mô hình biến thiên chậm của sao. Việc này được thực hiện bằng cách tối thiểu hóa một hàm mục tiêu $L(s)$, đo lường sai số còn lại giữa tín hiệu đã hiệu chỉnh và mô hình khử xu hướng.
Cụ thể, hàm mục tiêu được định nghĩa là sai số tuyệt đối trung bình (Mean Absolute Error - MAE):

$$L(s) = \frac{1}{N} \sum_{i=1}^{N} |F'(t_i, s) - P(t_i)|$$

Trong đó:

· $F'(t_i, s)$ là tín hiệu quang thông đã hiệu chỉnh tại thời điểm $t_i$,

· $P(t_i)$ là giá trị từ mô hình đa thức khử xu hướng tại thời điểm đó,

· $N$ là tổng số điểm dữ liệu.

Giá trị tối ưu của white light transit depth được xác định bằng:

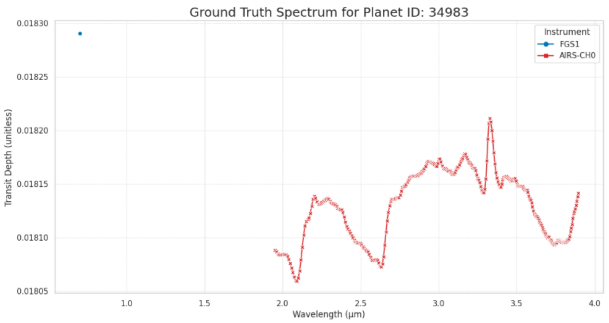
$$s_{\text{opt}} = \arg\min_{s} L(s)$$

### *3.3 Áp dụng mô hình Residual Network*

• Residual Network (ResNet) là một kiến trúc mạng nơ-ron sâu nổi bật với khả năng huấn luyện hiệu quả các mạng rất sâu nhờ cơ chế *residual learning*. Thay vì học trực tiếp một ánh xạ đầu vào $x \to H(x)$, ResNet học một hàm dư $F(x) = H(x) - x$, và kết hợp lại theo công thức:

$$H(x) = F(x) + x$$

- Bài toán yêu cầu dự đoán 283 giá trị của phổ truyền qua (transit spectrum), một đại lượng vật lý quan trọng để mô tả khí quyển của ngoại hành tinh. Trong quá trình xử lý, tôi nhận thấy giá trị *white light transit depth* (tức độ sâu quá cảnh trung bình trên toàn bộ dải sóng) đã là một ước lượng rất tốt cho toàn bộ phổ. Tuy nhiên, tại bước sóng $0.7\,\mu\text{m}$, giá trị này có sai số rất lớn so với tại các bước sóng khác. Đồng thời, kết quả ở bước sóng này đóng góp rất lớn vào số điểm cuối cùng với trọng số 0.4. Đây là điểm cần cải thiện bằng mô hình học sâu.




Hình 2. Giá trị phổ của hành tinh 34983 trong tập dữ liệu train, có thể thấy rằng giá trị tại bước sóng $0.7\,\mu\text{m}$ chênh lệch lớn so với giá trị của các bước sóng còn lại dẫn tới việc chênh lệch lớn với white light transit depth

Nhiệm vụ chính của mô hình học cách tinh chỉnh giá trị white transit depth. Nói cách khác, mô hình cần học phần dư (sự biến thiên của phổ tại bước sóng $0.7\,\mu\text{m}$) so với giá trị trung bình. Ý tưởng này có thể được biểu diễn bằng công thức:

$$\text{Spectrum}(\lambda) = \text{WhiteTransitDepth} + \Delta(\lambda, StarInfo) \tag{1}$$

Trong đó, nhiệm vụ của mạng nơ-ron là học hàm hiệu chỉnh $\Delta$, một hàm phức tạp phụ thuộc vào bước sóng $\lambda$ cùng các đặc tính của hành tinh và ngôi sao chủ $StarInfo$. Với bản chất này của bài toán, kiến trúc Mạng Residual (ResNet) là một lựa chọn tự nhiên và tối ưu. Các kết nối tắt (skip connections) trong ResNet cho phép mô hình truyền thẳng tín hiệu đầu vào khi không cần thay đổi, đồng thời chỉ tập trung học phần hiệu chỉnh $\Delta$.

### *3.4 Kiến trúc mạng*

*1) Lớp Đầu vào (Input Layer):* Dữ liệu đầu vào của mô hình bao gồm các thông tin:

- White Transit Depth: Một giá trị vô hướng.
- Đặc trưng hành tinh và sao chủ: Chỉ sử dụng 2 đặc trưng có tương tác mạnh nhất đến kết quả đầu ra, bao gồm: Ms (khối lượng sao được đo theo đơn vị khối lượng Mặt Trời, $M_\odot$) và $i$ (độ nghiêng quỹ đạo được đo theo đơn vị độ).

*Các Khối Residual (Residual Blocks)*

Phần lõi của mô hình là một chuỗi gồm 80 khối residual xếp chồng lên nhau, được định nghĩa trong lớp ResidualBlock. Mỗi khối xử lý một vector có chiều là 32. Cấu trúc của một khối được mô tả trong Bảng I.

Một điểm quan trọng là kết nối tắt (skip connection): đầu vào $x$ của khối (identity) được cộng trực tiếp với kết quả sau lớp BatchNorm1d thứ hai. Toàn bộ tổng này sau đó được đưa qua một hàm kích hoạt ReLU. Cấu trúc chi tiết có thể được biểu diễn như sau:

Bảng I
CẤU TRÚC CHI TIẾT CỦA MỘT KHỐI RESIDUAL DỰA TRÊN MÃ NGUỒN.

| Thứ tự | Lớp | Chi tiết / Kích hoạt |
|---|---|---|
| 1 | Linear 1 (fc1) | 32 đơn vị (units) |
| 2 | Batch Normalization 1 (bn1) | - |
| 3 | ReLU | Hàm kích hoạt |
| 4 | Dropout | Tỷ lệ: 0.3 |
| 5 | Linear 2 (fc2) | 32 đơn vị (units) |
| 6 | Batch Normalization 2 (bn2) | - |
| 7 | (Kết nối tắt) + ReLU | Cộng với đầu vào và qua ReLU |

*2) Lớp Đầu ra (Output Layer):* Lớp cuối cùng của mạng chứa 1 giá trị, tương ứng với giá trị tại bước sóng $0.7\,\mu\text{m}$.

*3) Hàm mất mát:*

- Hàm mất mát (Loss Function): Sử dụng hàm mất mát Mean Absolute Error (MAE) làm hàm mất mát trong quá trình huấn luyện. Công thức được định nghĩa như sau:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2}$$

Trong đó:

  - $N$ là tổng số mẫu trong tập dữ liệu.
  - $y_i$ là giá trị thực tế của mẫu thứ $i$.
  - $\hat{y}_i$ là giá trị do mô hình dự đoán cho mẫu thứ $i$.

## IV. THỰC NGHIỆM

### *4.1 Thiết lập*

Mô hình được huấn luyện và đánh giá trên tập dữ liệu của Ariel Data Challenge 2025, với dữ liệu được chia thành 80% mẫu huấn luyện và 20% mẫu kiểm tra (validation) theo dạng phân tầng để đảm bảo phân bố các đặc trưng hành tinh và ngôi sao chủ tương đồng ở cả hai tập.

Quy trình tiền xử lý dữ liệu áp dụng các bước đã trình bày. Các siêu tham số huấn luyện được chọn như sau:

- **Số epoch:** 200
- **Batch size:** 256
- **Learning rate:** $1 \times 10^{-3}$, điều chỉnh tự động bằng `ReduceLROnPlateau` với hệ số giảm 0.5, patience = 10 epoch
- **Optimizer:** AdamW với weight decay = $1 \times 10^{-4}$
- **Dropout:** 0.3 trong các khối residual
- **Seed:** 42 để đảm bảo tái lập kết quả

Môi trường thực nghiệm:

- **Framework:** PyTorch 2.2

### 4.2 Kết quả và phân tích

Bảng II trình bày kết quả so sánh giữa mô hình đề xuất và các biến thể ablation. Kết quả cho thấy:

- Sử dụng White Transit Depth đơn thuần đã cho kết quả tốt trên hầu hết các bước sóng, nhưng sai số vẫn cao tại $0.7\,\mu\mathrm{m}$.
- Mạng ResNet với 80 khối residual giúp giảm đáng kể MAE tại $0.7\,\mu\mathrm{m}$ nhờ khả năng học phần dư $\Delta$.
- Tăng số khối residual từ 20 lên 80 mang lại cải thiện rõ rệt, nhưng chi phí tính toán tăng khoảng $2.4\times$.
- Khi tăng lên 120 khối residual, MAE trên tập validation lại tăng, cho thấy mô hình bắt đầu bị overfitting do độ phức tạp vượt quá nhu cầu của dữ liệu.
- So với baseline, ResNet-80 giảm MAE 40,3%, trong khi ResNet-120 lại tăng nhẹ MAE do hiện tượng overfitting.

Bảng II
KẾT QUẢ TRÊN TẬP VALIDATION (MAE) TẠI $0.7\,\mu\mathrm{m}$

| **Mô hình** | **MAE**$_{0.7\mu m}$ |
|---|---|
| Baseline (White Transit Depth) | 0.001325 |
| ResNet (20 blocks) | 0.000950 |
| ResNet (80 blocks) | 0.000583 |
| ResNet (120 blocks) | 0.000712 |

## V. Ước lượng Độ không chắc chắn (Uncertainty Estimation)

Độ bất định (uncertainty) trong bài toán này được tính bằng độ lệch chuẩn của sai số dự đoán trên tập validation, phản ánh mức độ phân tán của các sai số quanh giá trị trung bình.

Gọi $y_i$ là giá trị thực và $\hat{y}_i$ là giá trị dự đoán tại mẫu thứ $i$, sai số được xác định bởi:

$$e_i = y_i - \hat{y}_i. \tag{3}$$

Độ bất định $U$ được tính như sau:

$$U = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2}, \tag{4}$$

trong đó $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$ là sai số trung bình và $n$ là số lượng mẫu trong tập validation. Công thức này sử dụng mẫu số $n$ do tập validation được coi là toàn bộ tập dữ liệu quan tâm. Từ công thức trên, độ không chắc chắn của mô hình ResNet được sử dụng để dự đoán bước sóng $0.7\,\mu\mathrm{m}$ là 0.00078, độ không chắc chắn của cách ước lượng giá trị 282 bước sóng còn lại là 0.0009.

## VI. Kết quả

Lời giải đạt 0.326 điểm trên bảng xếp hạng của cuộc thi.

## VII. Kết luận

Bài báo đã trình bày quy trình xây dựng mô hình học sâu dự đoán phổ quá cảnh từ dữ liệu mô phỏng của Ariel Data Challenge 2025. Phương pháp kết hợp giá trị *white light transit depth* với kiến trúc Residual Network cho phép mô hình học phần dư $\Delta$ để tinh chỉnh dự đoán tại các bước sóng $0.7\,\mu\mathrm{m}$.

Kết quả thực nghiệm cho thấy mô hình ResNet-80 đạt MAE thấp hơn 40,3% so với phương pháp baseline tại các bước sóng $0.7\,\mu\mathrm{m}$. Phân tích cho thấy việc tăng số khối residual trên mức tối ưu có thể dẫn đến hiện tượng overfitting.

## Tài liệu

[1] V. Kudelya, “Neurips: Non-ml transit curve fitting,” Kaggle, 2025, accessed: Aug. 10, 2025. [Online]. Available: https://www.kaggle.com/code/vitalykudelya/neurips-non-ml-transit-curve-fitting