Ariel Data Challenge 2025

A simplified neural-network approach

Luis Sauren

July 30, 2026

Problem

- Goal: reconstruct exoplanet transmission spectra from simulated Ariel observations.
- Ariel observes planets while they transit in front of their host stars.
- During the transit, the observed stellar flux decreases slightly.
- This decrease depends on wavelength and contains information about the planet atmosphere.

$$\text{noisy spectrometer data} \quad \longrightarrow \quad \text{transmission spectrum}$$

Dataset

- Input: AIRS-CH0 spectrometer files

$$\texttt{AIRS-CH0\_signal\_0.parquet}$$

- Raw shape:

$$(11250, 11392) = (11250, 32 \cdot 356)$$

- Reshaped to:

$$(11250, 32, 356)$$

- Target from `train.csv`:

$$y \in \mathbb{R}^{282}$$

We use `wl_2` to `wl_283`, because `wl_1` belongs to FGS1.

Simplifications

- The full Kaggle dataset is very large.
- No complete, optimized solution to the challenge
- Main simplifications:
  - only AIRS-CH0, no FGS1,
  - no calibration files,
  - only `signal_0`,
  - subset of 131 planets,
  - feature extraction before applying neural networks.

Aim: build a clear and understandable ML pipeline.

Preprocessing I: From frames to lightcurves

- ADC correction:

$$x_{\mathrm{corr}} = \frac{x_{\mathrm{raw}}}{\mathrm{gain}} + \mathrm{offset}.$$

- Reshape:

$$(11250, 11392) \to (11250, 32, 356).$$

- Frames are stored in low/high charge pairs.
- Flux is estimated by:

$$x_{\mathrm{flux}} = x_{1::2} - x_{0::2}.$$

- This gives:

$$(11250, 32, 356) \to (5625, 32, 356).$$

Preprocessing II: Transit-depth features

- Average over the 32 spatial pixels:

$$(5625, 32, 356) \to (5625, 356).$$

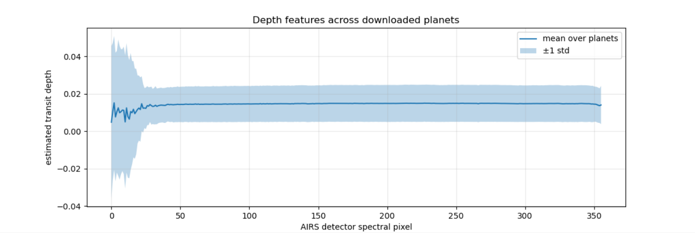
- Estimate transit depth for each spectral channel:

$$\mathrm{depth} = 1 - \frac{F_{\mathrm{in}}}{F_{\mathrm{out}}}.$$

- Assumption:
  - beginning and end: outside transit,
  - middle part: during transit.
- Use median flux values for robustness.

$$X \in \mathbb{R}^{356}$$

Feature cleaning and binning



- The first spectral channels show much larger fluctuations.
- Keep only channels 40 to 350:

$$X_{\mathrm{clean}} \in \mathbb{R}^{310}.$$

- Bin neighbouring channels and add global mean:

$$X_{\mathrm{reduced}} \in \mathbb{R}^{32}.$$

Models

Task:

$$X_{\text{reduced}} \in \mathbb{R}^{32} \longrightarrow y \in \mathbb{R}^{282}$$

Fully Connected NN

$$32 \to 64 \to 64 \to 282$$

1D-CNN

$$(N, 1, 32) \to \text{Conv1D} \to \text{Conv1D} \to \text{AvgPool} \to \text{FC} \to 282$$

The CNN uses shared filters along the spectral direction.

Training

- Train/validation split: 80/20.
- Inputs and targets standardized using training-set statistics.
- Loss function:

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{n}\sum_{i}(y_i - \hat{y}_i)^2.$$

- Optimizer: Adam.
- Learning rate:

$$\eta = 10^{-3}.$$

- Weight decay:

$$10^{-4}$$

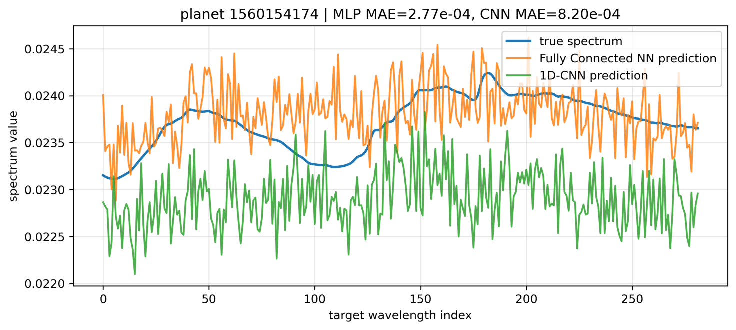
  as L2 regularization.
- Early stopping with patience 40.

Results

Model	MSE	MAE
Fully Connected NN	$2.0 \cdot 10^{-6}$	$9.88 \cdot 10^{-4}$
1D-CNN	$2.0 \cdot 10^{-6}$	$1.018 \cdot 10^{-3}$

- Both models reach similar performance.
- Fully connected network is slightly better in this run.
- CNN does not clearly outperform the simpler model.
- Likely reason: the input was already strongly compressed and binned.

Example prediction



Conclusion

- Built a simplified end-to-end pipeline for AIRS-CH0 data.
- Extracted physically motivated transit-depth features.
- Compared two neural-network architectures from the course.
- Both models produced reasonable spectra.
- Main limitations:
  - no calibration files,
  - no FGS1,
  - limited number of planets,
  - strong feature compression.