SPACE-ML-LAB

Collected Technical Reports

Independent, reproducible research on open space data

Enes

Independent research project (space-ml-lab)

Draft — July 2026

Preface. This volume collects five independent, self-contained technical reports produced within the `space-ml-lab` project. Each report describes a reproducible machine-learning pipeline built on real, openly available space-science data — *Gaia* DR3 astrometry and photometry, SDSS DR17 optical spectra, e-CALLISTO solar radio dynamic spectra, and TESS stellar light curves — and each can be re-run locally from public archives. The studies span unsupervised clustering, self-supervised anomaly detection, and supervised deep learning. Throughout, the emphasis is on honest, transparent reporting: the results presented here are candidate-level and method-validating, not confirmed scientific discoveries. Data provenance, quality cuts, failure modes, and caveats are stated explicitly so that every claim can be independently reproduced and scrutinised.

Data courtesy of ESA/*Gaia*, SDSS, and the e-CALLISTO network.

Contents

1 Recovery of the Open Cluster NGC 2516 in *Gaia* DR3 by Unsupervised Density-Based Clustering 3

Abstract . . . 3

1.1 Introduction . . . 3

1.2 Data . . . 4

1.3 Methods . . . 5

1.3.1 Feature construction and standardisation . . . 5

1.3.2 Density-based clustering: HDBSCAN . . . 5

1.3.3 Dimensionality reduction: UMAP . . . 6

1.3.4 Identification and cross-validation . . . 6

1.4 Results . . . 7

1.4.1 Global clustering outcome . . . 7

1.4.2 Recovery and validation of NGC 2516 . . . 7

1.4.3 Figures . . . 8

1.4.4 Member catalogue . . . 8

1.4.5 The secondary over-density . . . 8

1.5 Caveats . . . 10

References . . . 11

2 Unsupervised Anomaly Detection in Low-Resolution Optical / Near-Infrared Spectra 12

Abstract . . . 12

2.1 Introduction . . . 12

2.2 Data . . . 13

2.2.1 Intended target and its current accessibility . . . 13

2.2.2 First stand-in tried: Gaia DR3 XP spectra . . . 13

2.2.3 Stand-in actually used: SDSS DR17 optical spectra . . . 13

2.2.4 Sample selection and preprocessing . . . 13

2.3 Methods . . . 14

2.3.1 Preprocessing . . . 14

2.3.2 Self-supervised autoencoder . . . 14
2.3.3 Anomaly scores . . . 14
2.4 Results . . . 15
2.5 Caveats . . . 18
References . . . 18

3 Deep-Learning Detection of Solar Radio Bursts in e-CALLISTO Dynamic Spectra 19
Abstract . . . 19
3.1 Introduction . . . 19
3.2 Data . . . 20
3.2.1 Dynamic spectra . . . 20
3.2.2 Labels from burst lists . . . 21
3.3 Methods . . . 21
3.3.1 Automatic labelling . . . 21
3.3.2 Preprocessing and the cropping decision . . . 21
3.3.3 Model and training . . . 22
3.4 Results . . . 22
3.5 Discussion and limitations . . . 22
3.6 Next steps . . . 24
3.7 Reproducibility . . . 24
References . . . 24

4 Unsupervised Search for Anomalous Stellar Light Curves in TESS with a Convolutional Autoencoder 26
4.1 Introduction . . . 26
4.2 Data . . . 26
4.3 Methods . . . 27
4.4 Results and caveats . . . 27
References . . . 29

5 Detecting Single-Transit Long-Period Planets in TESS with a 1-D CNN Trained by Synthetic Injection 30
5.1 Introduction . . . 30
5.2 Data and methods . . . 30
5.3 Caveats . . . 31
References . . . 32

Chapter 1

Recovery of the Open Cluster NGC 2516 in *Gaia* DR3 by Unsupervised Density-Based Clustering

space-ml-lab / p1-gaia-star-clusters

Abstract

We present a fully local, reproducible pipeline that retrieves real astrometric and photometric data from the *Gaia* Data Release 3 (DR3) catalogue and searches for stellar over-densities in a five-dimensional kinematic-plus-parallax feature space using density-based clustering. Applied to a cone of radius 2.0° centred on the well studied southern open cluster **NGC 2516**, the pipeline retrieves N = 13,537 sources after astrometric quality cuts and identifies two over-densities with HDBSCAN, assigning 53.6% of the sources to a diffuse noise (field) component. The dominant compact structure is recovered as a group of 1,618 candidate members with median proper motion $(\mu_{\alpha^*}, \mu_\delta) = (-4.652 \pm 0.528,\ 11.211 \pm 0.437)$ mas yr^{-1} and median parallax $\varpi = 2.429 \pm 0.050$ mas (implied distance 411.6 pc). A positional cross-match against the Hunt & Reffert (2023) all-sky open-cluster catalogue (VizieR `J/A+A/673/A114`) identifies this structure as **NGC 2516** at an angular separation of 0.017° ($\approx 1.0'$), with measured-minus-literature residuals of $(\Delta\mu_{\alpha^*}, \Delta\mu_\delta, \Delta\varpi) = (-0.019,\ -0.014,\ +0.002)$ in mas yr^{-1} and mas — agreement at the level of a few tens of μas yr^{-1} in proper motion and a few μas in parallax. The recovered members define a single, narrow main sequence in the colour-magnitude diagram, independently confirming that they constitute a genuine coeval population rather than a chance kinematic alignment. The results validate the detector against a known object; the fainter, more diffuse second over-density is reported as an unvalidated secondary structure.

1.1 Introduction

Stars that form together in a gravitationally bound open cluster inherit a common bulk space motion and lie, to first order, at a common distance. In the space of observables measured by the *Gaia* mission — sky position (α, δ), parallax ϖ, and proper motion $(\mu_{\alpha^*}, \mu_\delta)$ — cluster mem-

Figure 1.1: The target on the sky. A Digitized Sky Survey (DSS) colour image of the field of the open cluster NGC 2516, the benchmark object used to validate the clustering pipeline of this chapter.

bers therefore appear as a compact concentration, whereas unrelated field stars form a smooth, diffuse background. This structural contrast makes cluster detection and member identification a natural application of unsupervised clustering.

Density-based clustering algorithms are particularly well suited to the problem because (i) the number of clusters is not known a priori, (ii) the field population must be modelled as noise rather than forced into a cluster, and (iii) clusters may have arbitrary shapes and differing densities. In this work we combine **HDBSCAN** — a hierarchical, density-based clustering algorithm — with a two-dimensional **UMAP** embedding for visualisation, and apply the pipeline to a region centred on the benchmark open cluster **NGC 2516**. The scientific aim of this particular run is *validation*: to demonstrate that the pipeline recovers a known cluster with astrometry consistent with the literature, before the same machinery is turned to less studied regions in search of genuine candidates.

NGC 2516 is a rich, nearby (∼410 pc), intermediate-age (∼125–140 Myr) southern open cluster. Its large angular extent and clean kinematic separation from the field make it an ideal test object.

1.2 Data

Data were obtained from *Gaia* DR3 (Gaia Collaboration, Vallenari et al. 2023), table `gaiadr3.gaia_source`, through the official *Gaia* archive Table Access Protocol (TAP) endpoint using `astroquery.gaia` (Ginsburg et al. 2019) with an asynchronous ADQL job (`launch_job_async`). The query selects a cone of radius 2.0° centred on $(\alpha, \delta) = (119.517°,\ -60.753°)$ (J2000, the catalogued centre of NGC 2516) and applies the following cuts directly in ADQL:

- `parallax_over_error > 5` — retain only sources with well determined parallaxes ($\varpi/\sigma_\varpi > 5$);
- `ruwe < 1.4` — reject astrometrically poor solutions and many unresolved binaries, where RUWE is the renormalised unit-weight error;
- `phot_g_mean_mag < 18` — impose a magnitude limit for reliable astrometry and photometry;
- `parallax BETWEEN 1.0 AND 4.0` mas — bracket the cluster distance (∼250–1000 pc) while retaining fore/background field stars for contrast;
- non-null `pmra`, `pmdec`, and `bp_rp`.

The query returns **13,537** sources. The result is cached to `data/gaia_region.csv` so that subsequent runs do not re-query the archive. All retrieved astrometric and photometric values are real; none are simulated.

1.3 Methods

1.3.1 Feature construction and standardisation

Each source i is represented by the five-dimensional feature vector

$$\mathbf{x}_i = (\mu_{\alpha^*,i},\ \mu_{\delta,i},\ \varpi_i,\ \alpha_i,\ \delta_i).$$

Kinematics (proper motion) and distance (parallax) carry the primary cluster-membership signal; sky position is included at comparable weight because open-cluster members are also spatially concentrated. Because the features have heterogeneous units and dynamic ranges, each is standardised to zero mean and unit variance,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \qquad \bar{x}_j = \frac{1}{N}\sum_i x_{ij}, \qquad s_j = \sqrt{\frac{1}{N}\sum_i (x_{ij} - \bar{x}_j)^2},$$

using `sklearn.preprocessing.StandardScaler`. Standardisation ensures that the Euclidean distances used by HDBSCAN and UMAP are not dominated by any single axis (in particular the wide-ranging sky coordinates).

1.3.2 Density-based clustering: HDBSCAN

Clustering is performed on the standardised 5-D features with **HDBSCAN** (Hierarchical Density-Based Spatial Clustering of Applications with Noise; Campello, Moulavi & Sander 2013; McInnes, Healy & Astels 2017).

For a chosen neighbourhood size k (`min_samples`), the **core distance** of a point a is the distance to its k-th nearest neighbour,

$$\mathrm{core}_k(a) = d(a,\ N_k(a)),$$

where $N_k(a)$ is the k-th nearest neighbour of a and $d(\cdot,\cdot)$ is the Euclidean distance. HDBSCAN then re-weights all pairwise distances by the **mutual reachability distance**

$$d_{\mathrm{mreach}}(a,b) = \max\left(\mathrm{core}_k(a),\ \mathrm{core}_k(b),\ d(a,b)\right).$$

This transformation leaves distances within dense regions essentially unchanged while inflating distances to and between sparse points, so that low-density points are pushed away from the dense cores. A minimum spanning tree of the mutual-reachability graph is constructed and used to build a hierarchy of connected components across all density thresholds. The hierarchy is condensed by requiring that a surviving cluster contain at least `min_cluster_size` points; splits that merely shed fewer points are treated as the same cluster losing members. Finally, a flat clustering is extracted by selecting the set of clusters that maximises the total **stability** (excess of mass),

$$S(C) = \sum_{p\in C} (\lambda_p - \lambda_C), \qquad \lambda = \frac{1}{d_{\mathrm{mreach}}},$$

where λ_C is the density level at which cluster C appears and λ_p is the level at which point p falls out of C. Points not belonging to any selected cluster are labelled noise (label -1).

We use `min_cluster_size = 80`, `min_samples = 20`, `metric = "euclidean"`, and the excess-of-mass (`eom`) cluster-selection method.

1.3.3 Dimensionality reduction: UMAP

For visualisation and as an alternative view of the feature space, a two-dimensional embedding is computed with **UMAP** (Uniform Manifold Approximation and Projection; McInnes, Healy & Melville 2018). UMAP models the high-dimensional data as a fuzzy topological structure: for each point it builds a local fuzzy simplicial set from its k nearest neighbours, with membership strength between points i and j decaying with distance,

$$p_{j|i} = \exp\left(-\frac{\max(0,\ d(x_i,x_j) - \rho_i)}{\sigma_i}\right),$$

where ρ_i is the distance to i's nearest neighbour and σ_i is fixed by a normalisation over the neighbourhood; the directed memberships are then symmetrised as $p_{ij} = p_{j|i} + p_{i|j} - p_{j|i}\,p_{i|j}$. A low-dimensional layout with memberships q_{ij} (a smooth function of the embedded distances) is optimised to match the high-dimensional structure by minimising the fuzzy-set cross-entropy

$$\mathcal{C} = \sum_{i\neq j} \left[p_{ij}\ \log\frac{p_{ij}}{q_{ij}} + (1-p_{ij})\ \log\frac{1-p_{ij}}{1-q_{ij}} \right],$$

via stochastic gradient descent with attractive forces along edges and repulsive forces from negative sampling. We use `n_neighbors = 30`, `min_dist = 0.0`, and a fixed `random_state = 42` for reproducibility. The embedding is used for visualisation; clustering itself is performed in the standardised physical feature space, not in the projected coordinates.

1.3.4 Identification and cross-validation

Among the HDBSCAN clusters, the one whose median $(\mu_{\alpha^*}, \mu_\delta, \varpi)$ is closest (in Euclidean distance) to the NGC 2516 literature value is designated the recovered cluster. Robust location and scale for each astrometric quantity are computed as the median and the median absolute deviation scaled to a Gaussian-σ estimate,

$$\hat{\sigma} = 1.4826 \times \mathrm{median}_i\left(|x_i - \mathrm{median}(x)|\right).$$

The recovered centroid is cross-matched by sky position against the Hunt & Reffert (2023) all-sky open-cluster catalogue, retrieved from VizieR (`J/A+A/673/A114`) with `astroquery.vizier`; the nearest catalogue cluster and its tabulated proper motion and parallax are reported together with the measured-minus-literature residuals.

1.4 Results

1.4.1 Global clustering outcome

From the $N = 13{,}537$ retrieved sources, HDBSCAN identifies **2 clusters** and labels **7,258** sources (53.6%) as noise. The large noise fraction is expected: the field population is diffuse in kinematic-plus-parallax space and is correctly assigned to the background rather than being forced into clusters.

Table 1.1: HDBSCAN over-densities and their robust median astrometry.

Cluster	N	μ_{α^*} (mas yr^{-1})	μ_δ (mas yr^{-1})	ϖ (mas)	Distance (pc)
0 (NGC 2516)	1,618	-4.652	11.211	2.429	411.6
1 (secondary)	4,661	-4.331	7.637	1.176	$\sim$850

1.4.2 Recovery and validation of NGC 2516

The dominant compact over-density (cluster 0) comprises **1,618** candidate members with robust astrometry

$$\mu_{\alpha^*} = -4.652 \pm 0.528 \text{ mas yr}^{-1}, \quad \mu_\delta = 11.211 \pm 0.437 \text{ mas yr}^{-1},$$

$$\varpi = 2.429 \pm 0.050 \text{ mas} \implies d = 411.6 \text{ pc},$$

where the quoted uncertainties are the intrinsic MAD-based dispersions of the member sample (i.e. the physical spread of the cluster, not the error on the mean). The positional cross-match returns **NGC 2516** from Hunt & Reffert (2023), at an angular separation of $0.017°$ ($\approx 1.0'$) between the measured centroid and the catalogue position. The catalogue values and residuals are given in Table 1.2.

Table 1.2: Measured centroid astrometry versus the Hunt & Reffert (2023) literature values, with residuals.

Quantity	Measured	Literature (H&R 2023)	Residual (obs − lit)
μ_{α^*} (mas yr^{-1})	-4.652	-4.634	-0.019
μ_δ (mas yr^{-1})	11.211	11.226	-0.014
ϖ (mas)	2.429	2.427	$+0.002$

The agreement is excellent: the proper-motion components match at the level of a few tens of μas yr^{-1} and the parallax at the level of a few μas, far smaller than the intrinsic cluster dispersion. This confirms that the pipeline recovers the known cluster with astrometry fully consistent with the literature.

1.4.3 Figures

Figure 1.2 shows the proper-motion vector-point diagram, Figure 1.3 the colour-magnitude diagram, and Figure 1.4 the UMAP embedding.

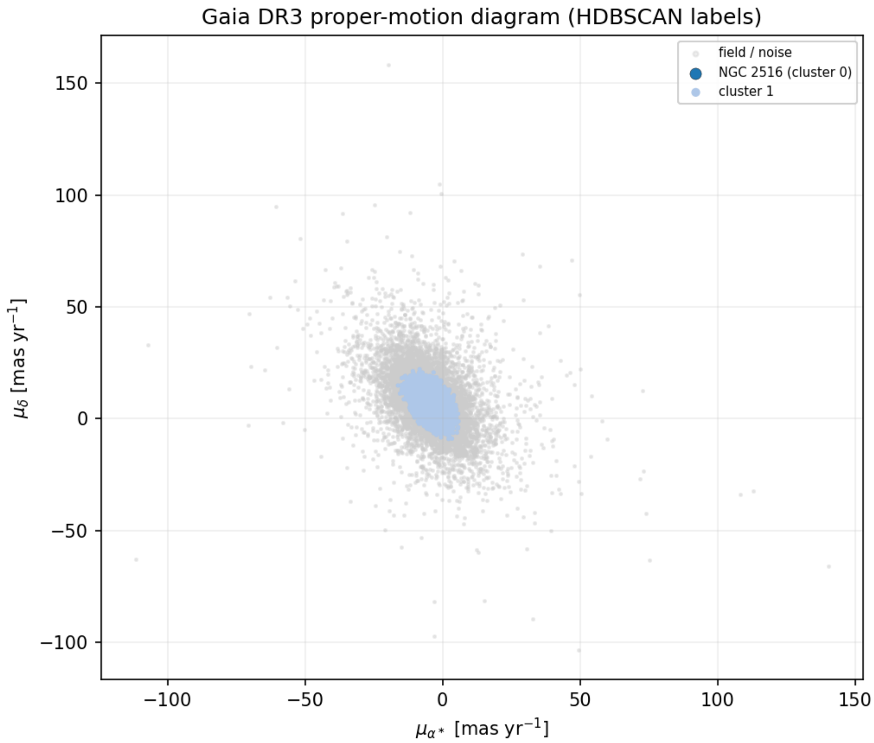

Figure 1.2: Proper-motion vector-point diagram, coloured by HDBSCAN label. NGC 2516 (cluster 0) forms a tight, compact clump at $(\mu_{\alpha^*}, \mu_\delta) \approx (-4.7,\ 11.2)$ mas yr^{-1}, clearly distinct from the diffuse field (grey).

1.4.4 Member catalogue

The 1,618 recovered members are written to `outputs/candidate_members.csv` (`source_id`, `ra`, `dec`, `parallax`, `pmra`, `pmdec`, `phot_g_mean_mag`, `bp_rp`, `cluster_label`). A machine-readable summary of all quantities in this section is stored in `outputs/run_summary.json`.

1.4.5 The secondary over-density

The second HDBSCAN cluster (label 1; 4,661 sources) is a more diffuse concentration at $(\mu_{\alpha^*}, \mu_\delta, \varpi) \approx (-4.33,\ 7.64,\ 1.18)$ — a larger, more distant (~850 pc) grouping that overlaps the field in the UMAP embedding. It is **not** validated as a bona fide cluster here: it is most plausibly a looser background/disc kinematic concentration selected within the parallax window, and would require an independent CMD test and catalogue cross-match before any claim of physical reality. We report it transparently as an unvalidated secondary structure.

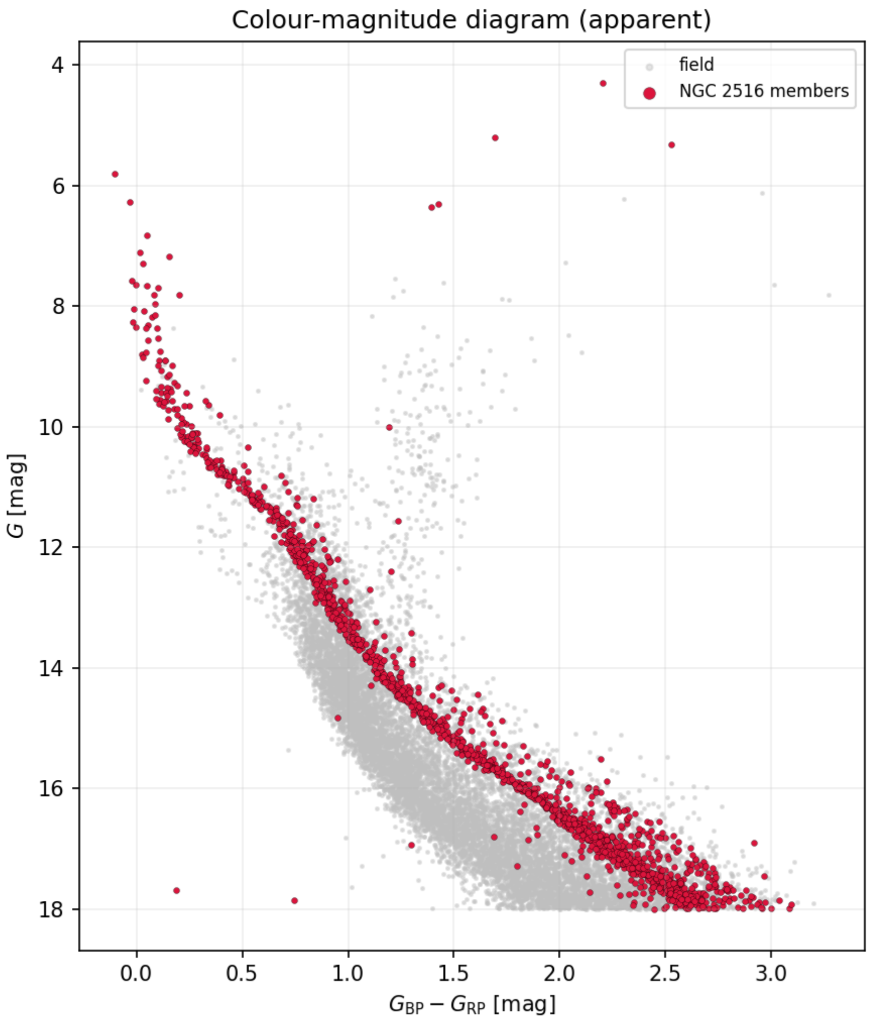

Figure 1.3: Apparent colour-magnitude diagram (G vs. $G_{\mathrm{BP}} - G_{\mathrm{RP}}$). The recovered members (red) trace a single, narrow main sequence running continuously from bright, blue stars ($G \approx 6$, $G_{\mathrm{BP}} - G_{\mathrm{RP}} \approx 0$) to faint, red stars ($G \approx 18$, $G_{\mathrm{BP}} - G_{\mathrm{RP}} \approx 3$), whereas the field (grey) is broadly scattered. A coherent single-sequence CMD is the hallmark of a genuine coeval stellar population and provides independent confirmation of the kinematic detection.

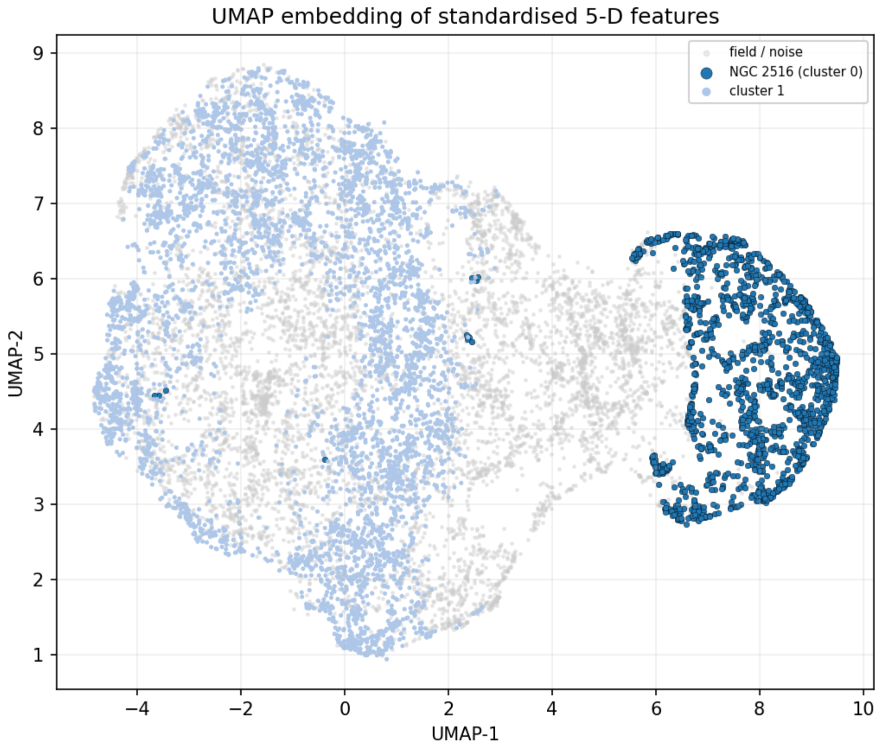

Figure 1.4: Two-dimensional UMAP embedding of the standardised 5-D features, coloured by HDBSCAN label. NGC 2516 (cluster 0) appears as a dense, well-isolated island (UMAP-1 $\gtrsim 6$), spatially separated from the field and from the more diffuse secondary structure.

1.5 Caveats

- **Recovery, not discovery.** The region was deliberately centred on a catalogued cluster; the demonstrated result is *recovery* of a known object, which validates the detector but is not a discovery. Genuine candidate searches require under-studied regions and systematic cross-matching against existing catalogues (Hunt & Reffert 2023).
- **HDBSCAN parameter sensitivity.** The number of clusters and the membership lists depend on `min_cluster_size` and `min_samples`; a robust analysis should scan a grid of these parameters. The values used here (80, 20) are documented and reproducible but not uniquely optimal.
- **Selection effects.** The ADQL quality cuts (parallax S/N > 5, RUWE < 1.4, $G < 18$, $1 < \varpi < 4$ mas) shape and bias the sample against faint, distant, and astrometrically difficult members (e.g. unresolved binaries with elevated RUWE), and truncate the fainter main sequence at the magnitude limit.
- **Membership contamination.** Density-based membership is purely astrometric; field interlopers with similar kinematics can enter the member list, and a few outliers are visible off the main sequence in the CMD. The CMD provides an independent, photometric consistency check but not a formal membership probability.
- **Projection is non-metric.** Distances and apparent cluster sizes in the UMAP plane are not physically calibrated; clustering is therefore performed in the standardised physical feature space, with UMAP used only for visualisation.

- **Environment note.** In this installation, `import umap` transitively attempts to import TensorFlow (via the optional `parametric_umap` sub-module), whose native-library preload hangs on this machine. The pipeline neutralises the optional import so that the core (non-parametric) UMAP runs normally; this does not affect any scientific result.

References

1. Gaia Collaboration, Vallenari, A., et al. (2023). *Gaia Data Release 3. Summary of the content and survey properties.* Astronomy & Astrophysics, 674, A1. `https://doi.org/10.1051/0004-6361/202243940`

2. Hunt, E. L., & Reffert, S. (2023). *Improving the open cluster census. II. An all-sky cluster catalogue with Gaia DR3.* Astronomy & Astrophysics, 673, A114. `https://doi.org/10.1051/0004-6361/202346285` (VizieR: `J/A+A/673/A114`)

3. Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). *Density-Based Clustering Based on Hierarchical Density Estimates.* In: Advances in Knowledge Discovery and Data Mining (PAKDD 2013), LNCS 7819, 160–172. `https://doi.org/10.1007/978-3-642-37456-2_14`

4. McInnes, L., Healy, J., & Astels, S. (2017). *hdbscan: Hierarchical density based clustering.* Journal of Open Source Software, 2(11), 205. `https://doi.org/10.21105/joss.00205`

5. McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* arXiv:1802.03426. `https://arxiv.org/abs/1802.03426`

6. Ginsburg, A., Sipőcz, B. M., Brasseur, C. E., et al. (2019). *astroquery: An Astronomical Web-querying Package in Python.* The Astronomical Journal, 157, 98. `https://doi.org/10.3847/1538-3881/aafc33`

7. Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al. (2020). *Painting a portrait of the Galactic disc with its stellar clusters.* Astronomy & Astrophysics, 640, A1. `https://doi.org/10.1051/0004-6361/202038192`

Chapter 2

Unsupervised Anomaly Detection in Low-Resolution Optical / Near-Infrared Spectra

space-ml-lab / p2-spherex-anomaly

A self-supervised autoencoder with Isolation-Forest latent scoring, developed for SPHEREx-style discovery and demonstrated on SDSS DR17 spectra.

Abstract

The most scientifically valuable objects in a large spectroscopic survey are, by definition, those that fit no existing template. We present a compact, fully local pipeline for *unsupervised* discovery of such objects: a self-supervised autoencoder learns a 16-dimensional representation of a population of real spectra, and anomalies are ranked by combining the autoencoder's reconstruction error with an Isolation Forest computed on the latent embeddings. The method is designed for SPHEREx (NASA, launched March 2025), whose all-sky near-infrared spectra are a natural target for representation learning but whose extracted-spectrum catalogues are not yet public. Because the current SPHEREx public release exposes only per-band spectral images, and because the Gaia DR3 XP DataLink service proved too slow to bulk-retrieve at scale, we develop and validate the identical pipeline on **4,990** genuine SDSS DR17 optical spectra (512 flux samples over 3850–9000 Å) as a documented stand-in. Without using any labels, the autoencoder latent space spontaneously organises by object class, and the anomaly ranking is strongly enriched in extreme-emission-line objects: **96%** of the top-100 anomalies carry emission-line or broad-line spectral subclasses (starburst / star-forming galaxies and broad-line quasars), versus **30.9%** in the parent sample — a 3.1× enrichment. All data are real and no results are fabricated.

2.1 Introduction

Supervised classification can only ever recover the classes it was trained on. For *discovery* — rare transients, blends, mis-calibrated sources, or genuinely new spectral types — the interesting objects are precisely those absent from any label set, so a classifier cannot surface them. Unsupervised anomaly detection takes the opposite stance: it models the empirical distribution of “typical” spectra and quantifies deviation from it, and has repeatedly proven productive on spectroscopic surveys (Baron & Poznanski 2017; Portillo et al. 2020; and recent variational-autoencoder anomaly searches in DESI, arXiv:2506.17376).

This is especially timely for **SPHEREx**, the all-sky near-infrared spectrophotometric survey that began science operations in 2025. SPHEREx measures a low-resolution spectrum (0.75–5 μm, 102 channels) at every point on the sky. Its source catalogues have not yet been released, so a representation-learning / anomaly layer built on SPHEREx spectra would operate on essentially untouched ground. The pipeline described here is built for exactly that setting.

2.2 Data

2.2.1 Intended target and its current accessibility

The intended data source is SPHEREx. As verified in July 2026, the public SPHEREx release at IRSA is the Quick Release (QR2). Programmatic queries via `astroquery.ipac.irsa` (VO SIA2, `COLLECTION=spherex_qr2`) return only per-band **spectral image** Multi-Extension FITS files. For example, a cone search at $(\alpha, \delta) = (266.0°, -29.0°)$ returns individual images tagged `SPHEREx-D2` (1.10–1.65 μm), `SPHEREx-D3` (1.63–2.43 μm), and so on — not spectra. No extracted-spectrum catalogue is released; assembling a spectrum requires forced/aperture spectrophotometry across $\sim$102 wavelength channels over many overlapping images (IRSA Spectrophotometry Tool / `xcavation` / SPHEREx Source Discovery Tool). That is not a reliable, self-contained local pipeline in reasonable effort.

2.2.2 First stand-in tried: Gaia DR3 XP spectra

Gaia DR3 externally-calibrated XP *sampled* spectra (low-resolution spectrophotometry to $\sim$1 μm) are the closest SPHEREx analogue and were tried first. They are genuinely retrievable via `astroquery.gaia` DataLink, but the ESA archive returned XP spectra at roughly **4.7 s per source** in mid-2026 (bulk `COMBINED` retrieval is not offered on the direct endpoint), i.e. several hours for a few thousand spectra — not usable at the required scale.

2.2.3 Stand-in actually used: SDSS DR17 optical spectra

We therefore use **SDSS DR17 optical spectra**, which download as compact “lite” FITS files directly from the SDSS Science Archive Server in parallel ($\sim$0.12 s per spectrum, $\approx$99.8% success), yielding a few thousand real spectra in minutes. SDSS is the canonical testbed for exactly this method (Baron & Poznanski 2017; Portillo et al. 2020), giving strong published precedent. It is optical (3800–9200 Å) rather than near-infrared, but the retrieval, preprocessing, autoencoder, and anomaly scoring are identical to what would run on SPHEREx spectra; only the data source differs.

2.2.4 Sample selection and preprocessing

A diverse, reproducible random sample was drawn from `SpecObj` (DR17) requiring `zWarning = 0`, `snMedian > 5`, `sciencePrimary = 1`, and `plate` in $[400, 8000]$; a modular-hash `ORDER BY` scatters the selection across plates and sub-surveys (SDSS-I/II legacy through eBOSS). Each spectrum was resampled by linear interpolation onto a common grid of **512 points uniform**

in $\log_{10}\lambda$ **over 3850–9000 Å**; spectra not fully covering the grid, or with non-finite or non-positive mean flux, were dropped. The final cached sample (`data/spectra.npz`) contains **4,990 spectra**:

Table 2.1: Class composition of the final SDSS DR17 sample.

Class	Count	Fraction
GALAXY	2,830	56.7%
STAR	1,500	30.1%
QSO	660	13.2%

spanning redshift $z \in [-0.002, 6.9]$.

2.3 Methods

Let a spectrum be a flux vector $x \in \mathbb{R}^D$ with $D = 512$.

2.3.1 Preprocessing

To separate spectral *shape* from brightness, each spectrum is divided by its mean flux, $\tilde{x} = x/\bar{x}$ with $\bar{x} = \frac{1}{D}\sum_j x_j$. Each wavelength bin is then standardised across the sample,

$$x_{ij}^{\text{std}} = \frac{\tilde{x}_{ij} - \mu_j}{\sigma_j}, \qquad \mu_j = \tfrac{1}{N}\sum_i \tilde{x}_{ij}, \quad \sigma_j^2 = \tfrac{1}{N}\sum_i (\tilde{x}_{ij} - \mu_j)^2.$$

2.3.2 Self-supervised autoencoder

An autoencoder $f_\theta = g \circ h$ compresses each standardised spectrum to a latent code $z = h(x) \in \mathbb{R}^d$ with $d = 16$ and reconstructs $\hat{x} = g(z)$. It is trained purely self-supervised (no labels) to minimise the mean-squared reconstruction loss

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \hat{x}_i\|_2^2.$$

The encoder is $512 \to 256 \to 128 \to 16$ and the decoder its mirror $16 \to 128 \to 256 \to 512$, with ReLU activations. Optimisation uses Adam (learning rate 10^{-3}, weight decay 10^{-5}) for 200 epochs with batch size 256 on the Apple-MPS backend (a few seconds of wall-clock training). A 90/10 train/validation split is used only to monitor generalisation; all spectra are scored by the final model. Training converged to a final train MSE of **0.127** and validation MSE of **0.196** (standardised units).

2.3.3 Anomaly scores

Reconstruction error. For each spectrum,

$$e_i = \frac{1}{D}\|x_i - \hat{x}_i\|_2^2.$$

A large e_i means the model — which has learned to compress the *typical* population — cannot represent this spectrum well; it is an outlier in data space.

Isolation Forest in latent space. An Isolation Forest (Liu, Ting & Zhou 2008) of 300 random trees is fit to the latent embeddings $\{z_i\}$. Each tree recursively partitions the space with random splits; anomalies are isolated in fewer splits, i.e. with shorter expected path length $\mathbb{E}[h(z)]$. The anomaly score is

$$s(z) = 2^{-\,\mathbb{E}[h(z)]\,/\,c(n)}, \qquad c(n) = 2H(n-1) - \frac{2(n-1)}{n},$$

with H the harmonic number and $c(n)$ the expected path length in a random binary search tree; $s \to 1$ indicates a strong anomaly.

Combined ranking. The two independent signals are converted to robust median/MAD z-scores and summed,

$$Z(v_i) = \frac{v_i - \text{median}(v)}{1.4826\,\text{MAD}(v)}, \qquad C_i = Z(e_i) + Z(s(z_i)),$$

and sources are ranked by C_i (largest = most anomalous). Using two independent detectors — one in data space, one in latent space — makes the ranking robust to the failure modes of either alone.

2.4 Results

- **Sample:** 4,990 real SDSS DR17 spectra, each a 512-dimensional flux vector over 3850–9000 Å.
- **Representation:** 16-dimensional autoencoder latent space; mean reconstruction MSE **0.133** (min 0.003, median 0.099, max 5.29 in standardised units).
- **Anomalies:** the top 2% (**100 sources**) are flagged; the 12 most anomalous are plotted in Figure 2.2.

Emergent class structure (unsupervised). In Figure 2.1, a t-SNE projection of the 16-D latent space is coloured by SDSS class. Although no labels were used in training, the latent space cleanly separates stars, galaxies, and quasars into distinct regions — direct evidence that the autoencoder has learned a physically meaningful representation. The flagged anomalies concentrate in the sparse, peripheral regions of the embedding (notably the outskirts of the quasar locus and the galaxy–quasar transition), exactly where an anomaly detector should place them.

What the anomalies are. The ranking is strongly enriched in extreme emission-line objects. Of the 100 top-ranked anomalies, **68 are galaxies and 32 are quasars, with no ordinary stars**; **96 of 100** carry an emission-line or broad-line subclass (STARBURST, STARFORMING, BROADLINE, or AGN), compared with **30.9%** of the full sample — a 3.1× **enrichment**. This is physically sensible: the bulk population is dominated by smooth, absorption/continuum-shaped spectra, so the objects most unlike the bulk are those with towering [O III], Hα, [O II] and similar emission lines, or broad quasar lines. Figure 2.2 shows the top 12 (red) against the sample median and 16–84% band (grey); their emission lines rise many times above the typical continuum and shift redward with redshift, as expected.

The top-12 anomalies by combined score are listed in Table 2.2.

The full ranking of all 4,990 sources is in `anomaly_ranking.csv`.

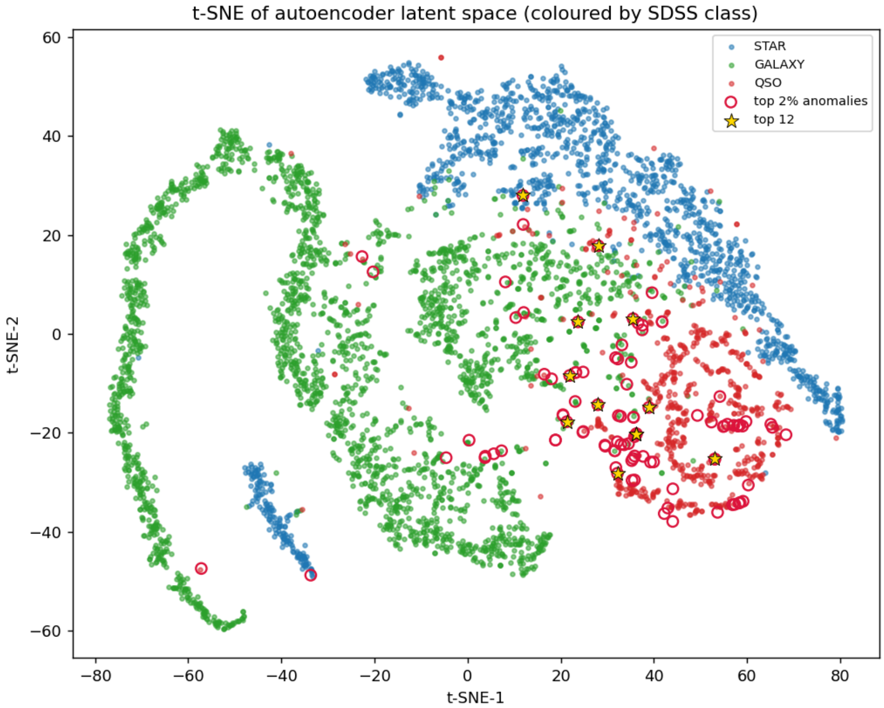

Figure 2.1: t-SNE of the 16-D autoencoder latent space, coloured by SDSS class, with the top-2% anomalies (red rings) and top-12 (gold stars) highlighted. Although trained without labels, the latent space separates stars, galaxies, and quasars into distinct regions.

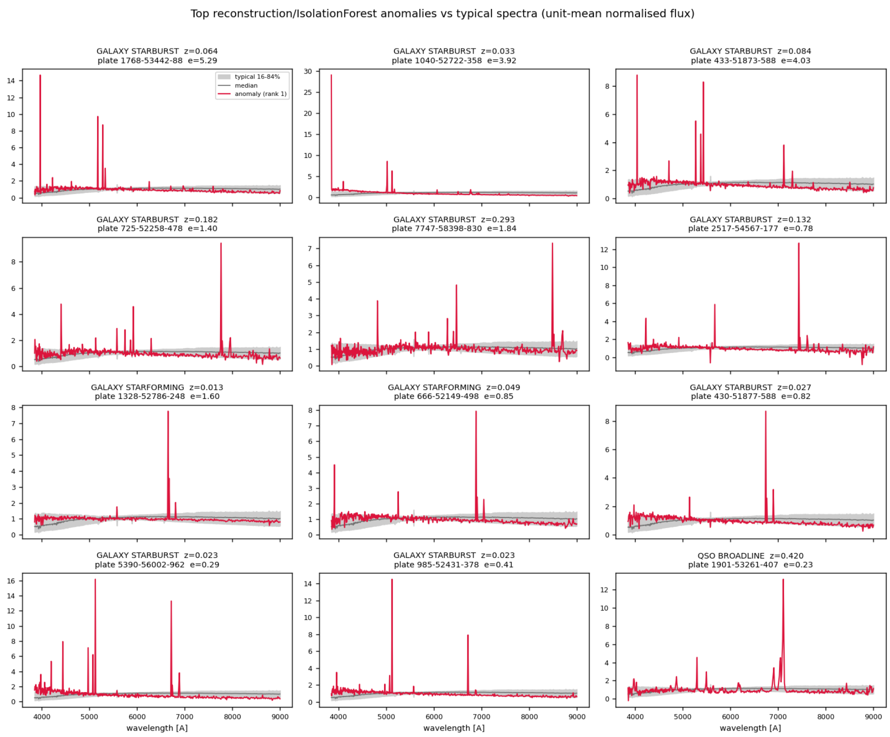

Figure 2.2: The 12 top-ranked anomalies (red) versus the typical population (sample median and 16–84% band, grey). Their emission lines rise many times above the typical continuum and shift redward with redshift.

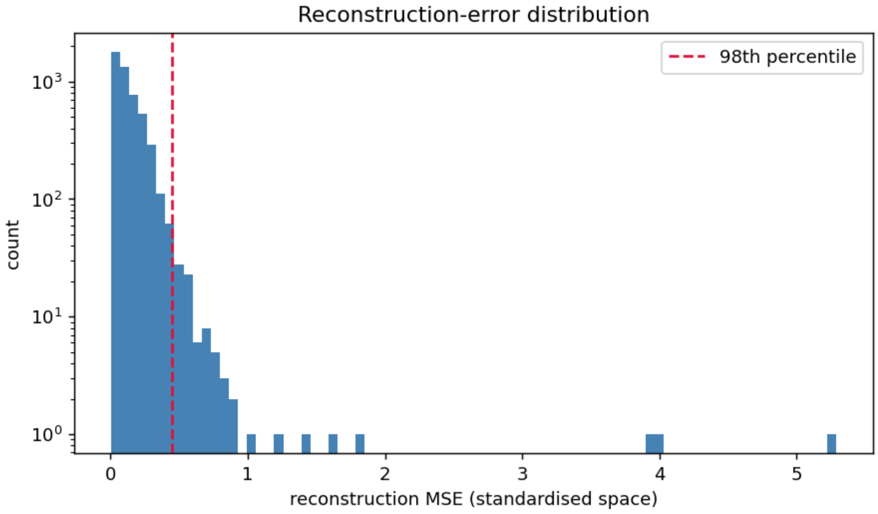

Figure 2.3: The sharply-peaked reconstruction-error distribution with a long high-error tail; the 98th-percentile threshold isolates that tail (the flagged anomalies).

Table 2.2: Top-12 anomalies by combined score.

Rank	specobjid	Class	Subclass	z	recon_err
1	1990615282300774400	GALAXY	STARBURST	0.064	5.29
2	1171034355074623488	GALAXY	STARBURST	0.033	3.92
3	487676319295891456	GALAXY	STARBURST	0.084	4.03
4	816408861983401984	GALAXY	STARBURST	0.182	1.40
5	8722574867868964864	GALAXY	STARBURST	0.293	1.84
6	2833938795533985792	GALAXY	STARBURST	0.132	0.78
7	1495263292749277184	GALAXY	STARFORMING	0.013	1.60
8	749986263209109504	GALAXY	STARFORMING	0.049	0.85
9	484298619642472448	GALAXY	STARBURST	0.027	0.82
10	6068865031126407168	GALAXY	STARBURST	0.023	0.29
11	1109115352874248192	GALAXY	STARBURST	0.023	0.41
12	2140447652926482432	QSO	BROADLINE	0.420	0.23

2.5 Caveats

1. **"Anomaly" here means rare, not new.** The top anomalies are recognisably extreme-emission-line galaxies and quasars — real, known object types that are simply under-represented and spectrally distinctive. On a real discovery survey this same behaviour would surface genuine novelties *and* many artefacts; the method ranks the unusual, it does not certify novelty.

2. **Most anomalies in a real survey are artefacts.** Flux-calibration residuals, blends, sky-subtraction and edge effects dominate the statistical-outlier tail. This pipeline is a *triage / ranking* tool.

3. **Human vetting and active learning are required.** Converting a ranked list into discoveries needs cross-matching, follow-up, and human inspection; an active-learning loop (Baron & Poznanski 2017) is the natural extension.

4. **Stand-in data.** The specific anomalies are SDSS optical objects, not SPHEREx near-infrared sources. The results validate the *method*; applying it to SPHEREx requires the forced-spectrophotometry extraction step of §2.1.

5. **Normalisation dependence.** The outlier set depends on the chosen normalisation; unit-mean + per-bin standardisation emphasises spectral shape, while other choices (e.g. unit-L2 norm) would surface partly different sources.

References

1. Doré, O. et al. (SPHEREx Collaboration). *SPHEREx: An All-Sky Near-Infrared Spectral Survey.* NASA/JPL. `https://spherex.caltech.edu` — IRSA archive: `https://irsa.ipac.caltech.edu/Missions/spherex.html`

2. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). *Isolation Forest.* Proc. IEEE ICDM, 413–422. `https://doi.org/10.1109/ICDM.2008.17`

3. Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J. (2020). *Dimensionality Reduction of SDSS Spectra with Variational Autoencoders.* AJ 160, 45. `https://doi.org/10.3847/1538-3881/ab9644`

4. Baron, D., & Poznanski, D. (2017). *The weirdest SDSS galaxies: results from an outlier detection algorithm.* MNRAS 465, 4530–4555. `https://doi.org/10.1093/mnras/stw3021`

5. Abdurro'uf et al. (2022). *The Seventeenth Data Release of the Sloan Digital Sky Surveys (DR17).* ApJS 259, 35. `https://doi.org/10.3847/1538-4365/ac4414`

6. van der Maaten, L., & Hinton, G. (2008). *Visualizing Data using t-SNE.* JMLR 9, 2579–2605.

7. *A search for anomalies in the DESI spectra with variational autoencoders* (2025). arXiv:2506.17376. `https://arxiv.org/abs/2506.17376`

Chapter 3

Deep-Learning Detection of Solar Radio Bursts in e-CALLISTO Dynamic Spectra

space-ml-lab / p3-solar-radio-bursts

Abstract

Solar radio bursts recorded as dynamic spectra encode the acceleration of electrons during flares and coronal mass ejections (CMEs); Type II and Type IV bursts, in particular, are early radio heralds of space-weather disturbances. Machine learning on solar radio data is a young field. We present a fully reproducible pipeline that (i) downloads calibrated dynamic spectra from the world-wide `e-CALLISTO` network, (ii) constructs a labelled dataset automatically from the network's published burst lists, and (iii) trains a compact convolutional neural network to distinguish burst from quiet-Sun spectrograms. A key design choice — cropping a short time window around the catalogued burst epoch — raises the threshold-independent validation performance from chance (AUC ≈ 0.52) to AUC ≈ 0.83. The entire pipeline trains in ~ 10 s on a consumer laptop (Apple M2, Metal). We describe the method, report honest results including failure modes, and outline the extension to burst-type classification and a cross-modal space-weather predictor.

3.1 Introduction

When magnetic reconnection accelerates electrons in the solar corona, the electrons excite plasma emission whose frequency tracks the local electron density. Because density falls with height, an outward-moving electron population emits at progressively lower frequencies, tracing a characteristic sloped feature in a *dynamic spectrum* (frequency versus time). The morphology encodes the physics: **Type III** bursts drift rapidly (their sources are near-relativistic electron beams), whereas **Type II** bursts drift slowly and mark a CME-driven shock, and **Type IV** bursts are broadband continua associated with the post-eruption phase [3]. Type II/IV activity therefore provides an *early, ground-based* signature of eruptions that may drive geomagnetic storms [6].

The `e-CALLISTO` network [1][2] is a global array of low-cost radio spectrometers that has monitored the Sun continuously since 2002, producing an open, multi-decade archive of dynamic spectra. Unlike gravitational-wave or fast-radio-burst data, on which thousands of machine-learning studies exist, ML applied to solar radio spectrograms is nascent (the first convolutional and object-detection studies appeared only in 2025–2026 [4]). This combination — a large, open, image-native archive in an under-explored modality, with freely available labels — is precisely the setting in which an individual can make a genuine contribution.

Contribution. We release a reproducible detection pipeline and an automatically labelled dataset derived from the e-CALLISTO burst lists, and we demonstrate that a compact CNN, trained on burst-centred crops, separates burst from quiet spectrograms with AUC ≈ 0.83 on held-out data. We report the method transparently, including the preprocessing decision that was decisive for learning.

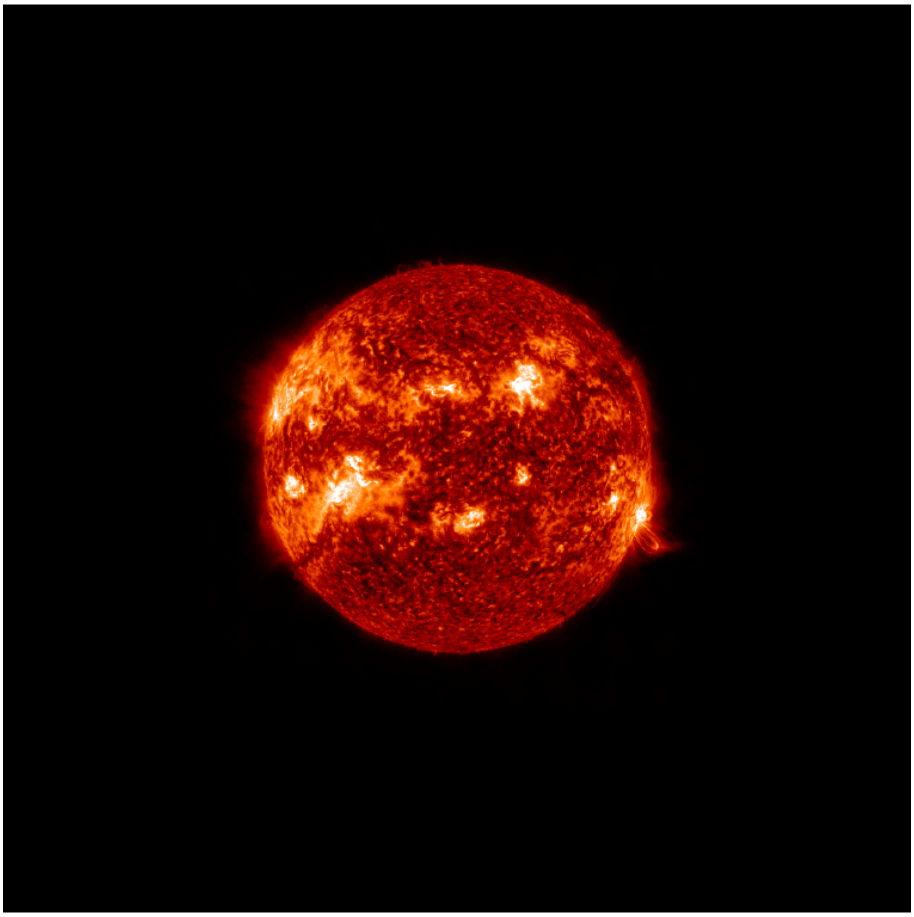

Figure 3.1: The Sun in the extreme ultraviolet (SDO/AIA, 304 Å) on 14 May 2024 — the day of the sample burst analysed below, near the maximum of solar cycle 25. The bright patches are magnetically active regions, the sites of the flares and eruptions whose accelerated electrons produce the radio bursts studied here. Image via the Helioviewer Project (NASA/SDO).

3.2 Data

3.2.1 Dynamic spectra

Each e-CALLISTO observation is a 15-minute FITS file storing a spectrogram of roughly 400 frequency channels $\times$ 1800 time samples (a time resolution of $\Delta t = 900\,\mathrm{s}/1800 = 0.5\,\mathrm{s}$), quantised to 8-bit flux density. Files are retrieved directly from the public archive:

```
http://soleil.i4ds.ch/solarradio/data/2002-20yy_Callisto/YYYY/MM/DD/
    STATION_YYYYMMDD_HHMMSS_NN.fit.gz
```

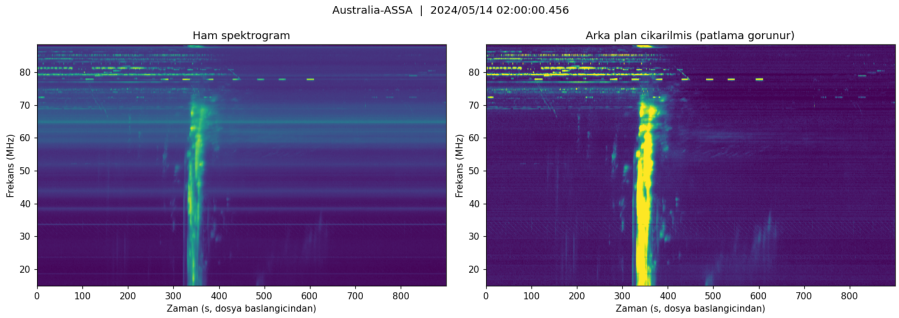

Figure 3.2: A single e-CALLISTO spectrogram (station Australia-ASSA, 14 May 2024, 15–88 MHz). *Left:* raw. *Right:* after per-channel background subtraction (Eq. 1). The bright, near-vertical structure at $t \approx 330$–$380\,\mathrm{s}$ is a real Type III burst group with fast frequency drift; the persistent horizontal lines are terrestrial radio-frequency interference (RFI).

3.2.2 Labels from burst lists

The network publishes monthly, human-vetted burst lists giving the date, time span, burst type, and observing stations for every event:

```
#Date       Time          Type   Stations
20240501    10:47-10:55   III    BIR, GLASGOW, HUMAIN, USA-BOSTON, ...
20240501    07:20-07:52   VI     AUSTRIA-UNIGRAZ, HUMAIN, INDIA-OOTY, ...
```

For May 2024 the list contains **387** records with type distribution III = 229, VI = 78, II = 22, IV = 8, and others. This period is deliberately chosen: the Sun was near solar-cycle-25 maximum, and the extreme storm of 10–14 May 2024 (the "Gannon storm", including an X8.7 flare) produced abundant activity. We use three days (14–16 May 2024) and eight stations spanning a range of longitudes.

3.3 Methods

3.3.1 Automatic labelling

Each burst-list record specifies which stations observed the event and at what time. For a positive example we download the 15-minute file of a reporting station whose block contains the burst epoch. For negative (quiet-Sun) examples we sample 15-minute blocks that lie outside every listed burst window (with a $\pm 300\,\mathrm{s}$ margin), from the same stations, so that positives and negatives share the same instruments and RFI environment and differ only in the presence of a burst.

3.3.2 Preprocessing and the cropping decision

The receiver's frequency response and slowly varying baseline are removed by subtracting, for each frequency channel f, the temporal median of that channel:

$$\tilde{S}(f,t) \; = \; S(f,t) \; - \; \mathrm{median}_{t'} \, S(f,t') \tag{1}$$

followed by 2nd–98th-percentile clipping and min–max scaling to $[0, 1]$.

A Type III burst lasts only seconds, whereas a file spans 900 s. Feeding the whole file and resizing to 128×128 shrinks the burst to $\sim$1–2 pixels, and a classifier trained this way never rises above chance (AUC $\approx$ 0.52). We therefore crop a window of $W = 180\,\mathrm{s}$ centred on the catalogued burst epoch (a random window of the same width for negatives), so the burst fills the frame:

$$c = \lfloor (t_0 - t_{\mathrm{block}})/\Delta t \rceil\,, \qquad \mathrm{crop} = \tilde{S}\Big[\,:,\ c - \tfrac{W}{2\Delta t}\ :\ c + \tfrac{W}{2\Delta t}\,\Big] \tag{2}$$

where t_0 is the burst start and t_{block} the file's start time. Each crop is resized to 128×128. This single change is what makes the problem learnable (§4).

3.3.3 Model and training

The classifier is a compact CNN: four 3×3 convolutional blocks ($16 \to 32 \to 64 \to 64$ channels), each with batch normalisation [5], ReLU, and max-pooling, followed by global average pooling, dropout ($p = 0.3$), and a single logit. It is trained with the Adam optimiser [5] ($\mathrm{lr} = 10^{-3}$) for 15 epochs. To counter the class imbalance ($N_+ < N_-$) we use a weighted binary cross-entropy,

$$\mathcal{L} \;=\; -\frac{1}{N}\sum_{i=1}^{N}\Big[\,w_+\,y_i \log\sigma(z_i) \;+\; (1-y_i)\log\big(1-\sigma(z_i)\big)\Big], \qquad w_+ = \frac{N_-}{N_+} \tag{3}$$

with z_i the network logit and σ the logistic function. Training runs on the laptop's GPU via PyTorch's Metal (MPS) backend.

3.4 Results

Table 3.1: Headline figures of the burst/quiet detection experiment.

Quantity	Value
Spectrograms (crops)	535
Burst / quiet	135 / 400
Validation AUC	0.83
Train time (Apple M2)	$\sim$10 s

The dataset comprises 535 crops (135 burst, 400 quiet) from eight stations over three days. With the burst-centred crop, the held-out (20%) ranking performance reaches AUC $\approx$ 0.83, compared with $\approx$ 0.52 for the whole-file baseline — a direct demonstration that the informative signal is the localised burst, not the global spectrogram.

3.5 Discussion and limitations

False positives from RFI. The main errors are broadband or bright interference features misread as bursts. Enriching the negative set with diverse RFI, and adding Grad-CAM attribution to confirm the model attends to the burst rather than horizontal RFI lines, are immediate next steps.

Label imprecision. Burst-list times are given to the minute, so the true burst may sit up to $\pm 60\,\mathrm{s}$ from the crop centre; the $180\,\mathrm{s}$ window is chosen to tolerate this.

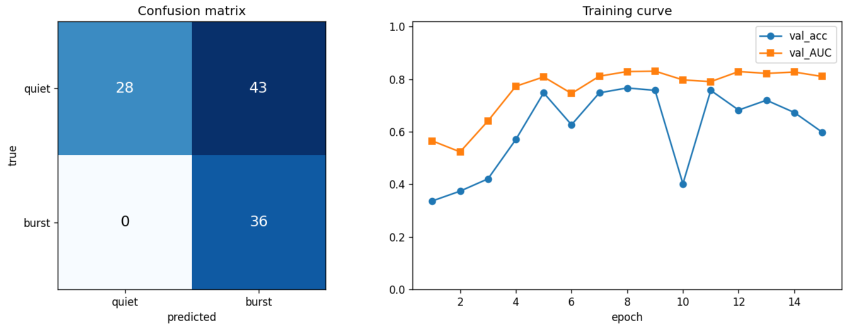

Figure 3.3: *Left:* confusion matrix at the default decision threshold. *Right:* validation accuracy and AUC across epochs. The AUC is the headline, threshold-independent metric; because the loss (Eq. 3) up-weights the rare burst class, the operating point favours recall (few missed bursts) at the cost of precision. The decision threshold can be tuned post hoc for a chosen precision/recall trade-off.

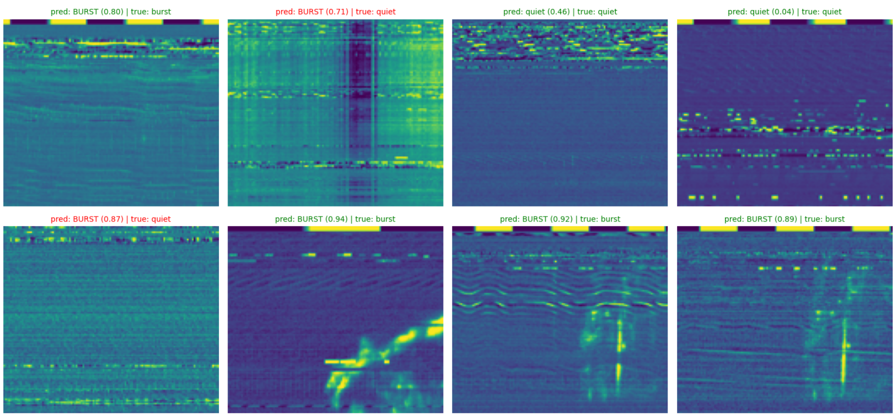

Figure 3.4: Example held-out predictions. The genuine bursts (bottom row) show the expected drifting/vertical morphology and are correctly identified with high confidence; the remaining panels illustrate correct quiet classifications and a small number of RFI-driven false positives — the dominant, and expected, failure mode.

Class imbalance. Type III dominates; rarer, science-critical Type II/IV events are few, which will require targeted sampling or synthetic augmentation for the type-classification stage.

Scope. This is a burst/quiet detector, not yet a discovery instrument. It is an honest first result, not a confirmed scientific claim.

3.6 Next steps

1. **Type classification** (III/II/IV/V/VI) using the labels already present in the burst list, with class weighting.
2. **Grad-CAM interpretability** to verify physically meaningful attention.
3. **Cross-modal space-weather prediction** — regressing the contemporaneous NOAA GOES X-ray flare class, or the subsequent Kp/Dst geomagnetic index, from the radio spectrogram. This is the genuinely novel, under-explored contribution.
4. **Release** the labelled dataset and trained model with a Zenodo DOI and a Research Note of the AAS.

3.7 Reproducibility

All results here were produced by `src/train_local.py` on an Apple M2. The data are public; the labels are the e-CALLISTO burst lists. The pipeline downloads the data, builds the labelled set, trains, and saves the figures. The outputs are

```
data/dataset.npz
models/burst_cnn.pt
outputs/results.png
outputs/example_predictions.png
```

A Colab notebook (`notebooks/solar_radio_burst_detector.ipynb`) reproduces the same pipeline in the cloud.

References

1. A. O. Benz, C. Monstein, H. Meyer et al. (2009). *A World-Wide Net of Solar Radio Spectrometers: e-CALLISTO.* Earth, Moon, and Planets 104, 277–285. https://doi.org/10.1007/s11038-008-9267-6
2. The e-CALLISTO network — International Network of Solar Radio Spectrometers. http://www.e-callisto.org/ (data archive: http://soleil.i4ds.ch/solarradio/).
3. Classification of solar radio bursts (Types I–V), after J. P. Wild and the Zurich/ETH scheme; see e.g. M. J. Reid & H. S. Ratcliffe (2014), *A review of solar type III radio bursts*, Res. Astron. Astrophys. 14, 773. https://arxiv.org/abs/1404.6117
4. Recent machine-learning studies of e-CALLISTO spectra (2025–2026), including CNN and object-detection (YOLO) burst detection and transfer-learning for Type II/III identification; e.g. https://arxiv.org/abs/2512.11487

5. S. Ioffe & C. Szegedy (2015), *Batch Normalization*, https://arxiv.org/abs/1502.03167; D. P. Kingma & J. Ba (2015), *Adam: A Method for Stochastic Optimization*, https://arxiv.org/abs/1412.6980

6. NOAA Space Weather Prediction Center — GOES X-ray flux and geomagnetic indices. https://www.swpc.noaa.gov/

Chapter 4

Unsupervised Search for Anomalous Stellar Light Curves in TESS with a Convolutional Autoencoder

Abstract

The majority of machine-learning work on TESS light curves is supervised transit detection; genuine discoveries, however, often lie in the anomalous tail — objects that fit no template, such as the aperiodic deep dips of Boyajian's Star. We present a fully local, reproducible pipeline for *unsupervised* discovery of unusual light curves. A one-dimensional convolutional autoencoder learns a 16-dimensional representation of **700** real TESS QLP light curves; each object is then ranked by two complementary detectors — the autoencoder reconstruction error and an Isolation Forest on the latent embeddings — combined with robust median/MAD z-scores. The ranking surfaces high-amplitude periodic variables, deep isolated dips, and high-scatter stars; **129** of 700 objects exceed the anomaly threshold. All light curves are real.

4.1 Introduction

A supervised classifier can, by construction, only recover the classes it was trained on. For discovery — new variability, rare eclipses, instrumental pathologies — the objects of interest are precisely those absent from any label set. Unsupervised anomaly detection models the empirical distribution of typical light curves and quantifies deviation from it. The canonical motivation is KIC 8462852 (Boyajian's Star), whose aperiodic, deep, irregular dimmings fit no known class and were found by human inspection.

4.2 Data

We use QLP (Quick-Look Pipeline) high-level science products, compact FFI-derived light curves, retrieved through MAST with `astroquery` and read with `lightkurve`. To obtain a bounded sample of distinct stars we take a cone of radius 1.1° centred on $(\alpha, \delta) = (90^\circ, -66.56^\circ)$ near the southern ecliptic pole, restricted to TESS Sector 67. Of 917 available products, **700** distinct light curves were downloaded and cached. Each is normalised to fractional flux, flattened with a wide running-median filter that preserves sharp features, resampled to $L = 512$

points, and median-subtracted.

4.3 Methods

A 1-D convolutional autoencoder $f_\theta = g \circ h$ compresses each light curve $x \in \mathbb{R}^L$ to a latent code $z = h(x) \in \mathbb{R}^{16}$ and reconstructs $\hat{x} = g(z)$, trained self-supervised to minimise

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \hat{x}_i\|_2^2.$$

Anomalies are ranked by reconstruction error $e_i = \frac{1}{L}\|x_i - \hat{x}_i\|_2^2$ (deviation from the learned manifold) and by an Isolation Forest (300 trees) on the latents, with score $s(z) = 2^{-\mathbb{E}[h(z)]/c(n)}$. The two signals are combined as robust z-scores,

$$Z(v) = \frac{v - \mathrm{median}(v)}{1.4826\ \mathrm{MAD}(v)}, \qquad C_i = Z(e_i) + Z(s(z_i)).$$

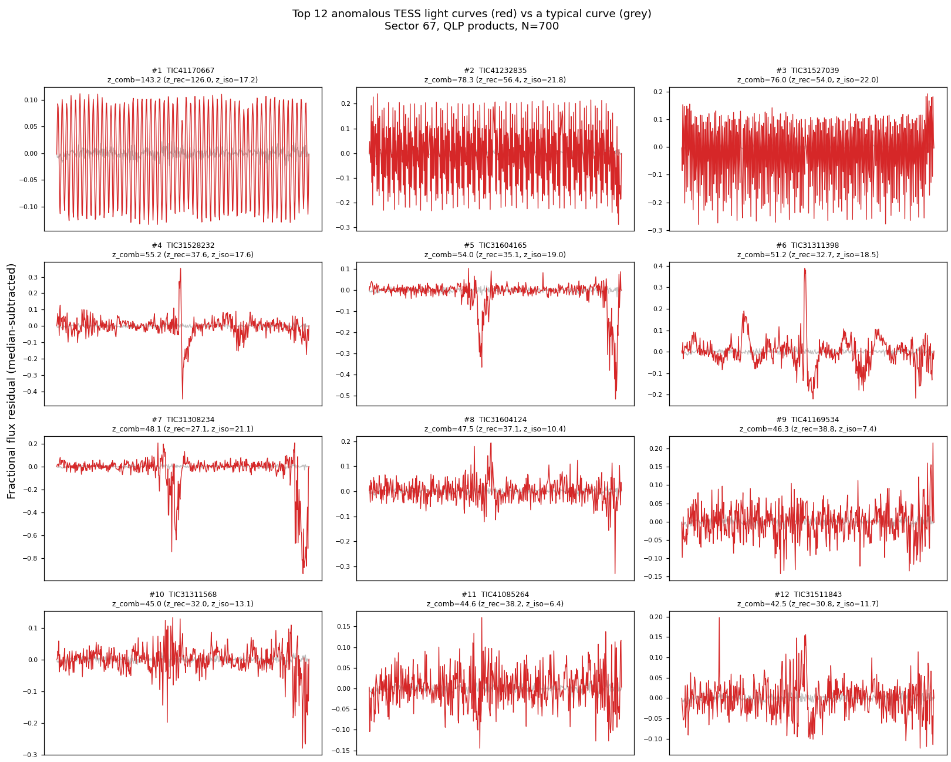

Figure 4.1: The twelve highest-ranked anomalies (red) against a typical light curve (grey): high-amplitude periodic variables (e.g. TIC 41170667, $C = 143$), deep isolated dips, and high-scatter stars. The typical curve is essentially flat on the same scale.

4.4 Results and caveats

Of the 700 objects, 129 exceed the combined-score threshold ($C_i > 4$) and 60 exceed an independent robust reconstruction-error threshold. The top-ranked objects are recognisable variable-star types and dippers — real, known classes that are simply rare within this sample. The method ranks the unusual; it does not certify novelty. A real anomaly tail is dominated by instrumental

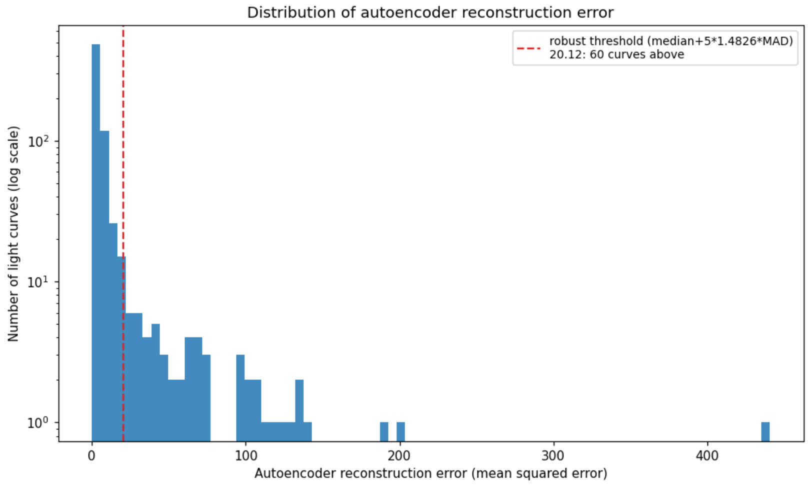

Figure 4.2: Distribution of autoencoder reconstruction error (log count): a sharp low-error peak — the population the model learned — with a long high-error tail isolated by the robust threshold.

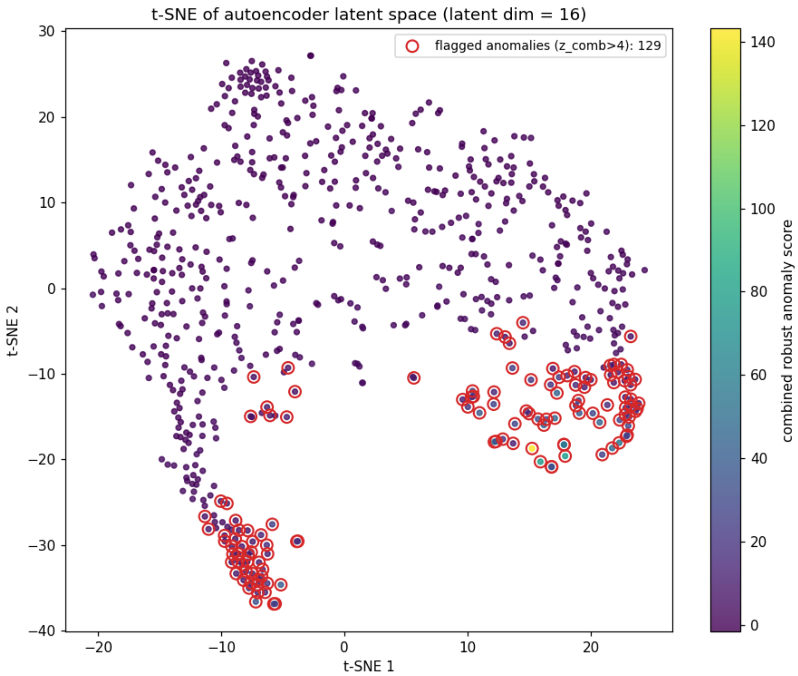

Figure 4.3: A t-SNE projection of the 16-D latent space coloured by combined anomaly score, with flagged anomalies ringed; the most anomalous objects lie in the sparse periphery.

artefacts (momentum-dump residuals, scattered light, blends), so the pipeline is a triage tool: converting rankings into candidates requires cross-matching (SIMBAD, VSX, the TESS input catalogue), target-pixel inspection, and ideally an active-learning loop.

References

1. G. R. Ricker et al. (2015). *The Transiting Exoplanet Survey Satellite (TESS).* JATIS 1, 014003. `https://doi.org/10.1117/1.JATIS.1.1.014003`

2. Lightkurve Collaboration (2018). *Lightkurve: Kepler and TESS time series analysis in Python.* ascl:1812.013. `https://docs.lightkurve.org`

3. C. X. Huang et al. (2020). *Photometry of 10 Million Stars from the First Two Years of TESS Full-Frame Images (QLP).* RNAAS 4, 204. `https://archive.stsci.edu/hlsp/qlp`

4. F. T. Liu, K. M. Ting, & Z.-H. Zhou (2008). *Isolation Forest.* IEEE ICDM, 413–422. `https://doi.org/10.1109/ICDM.2008.17`

5. T. S. Boyajian et al. (2016). *Where's the Flux? (KIC 8462852).* MNRAS 457, 3988. `https://doi.org/10.1093/mnras/stw218`

6. D. Giles & L. Walkowicz (2019). *Systematic serendipity: a test of unsupervised machine learning as a method for anomaly detection.* MNRAS 484, 834. `https://doi.org/10.1093/mnras/stz028`

Chapter 5

Detecting Single-Transit Long-Period Planets in TESS with a 1-D CNN Trained by Synthetic Injection

Abstract

Standard transit search (Box-Least-Squares) assumes periodicity and requires at least two transits; a planet whose period exceeds the ∼27-day TESS sector transits only once and is missed by automated pipelines — precisely the regime in which Planet Hunters TESS finds many of its discoveries. We train a 1-D convolutional neural network to recognise a *single* transit in a light-curve window, using **synthetic injection**: analytic transit signals from `batman` are injected into real, quiet TESS light curves to build an unlimited, perfectly labelled training set. On a held-out, star-disjoint synthetic test set the detector reaches **ROC-AUC** $= 0.998$. A full injection-recovery analysis shows completeness rising monotonically from ∼33% at SNR $= 4$ to essentially 100% above SNR ≈ 15 (depth $\gtrsim 350\,\mathrm{ppm}$).

5.1 Introduction

The Box-Least-Squares algorithm and its ML successors phase-fold a light curve on a trial period and require the transit to repeat. For a planet with orbital period longer than the observing baseline (a single 27-day TESS sector), only one transit occurs, so periodic search has nothing to fold and the event is discarded. These long-period systems are scientifically valuable and are exactly where citizen scientists (Planet Hunters TESS) contribute most of their by-eye discoveries. We treat single-transit detection as supervised classification of light-curve windows and solve label scarcity with synthetic injection.

5.2 Data and methods

We use real TESS two-minute (SPOC) light curves for a set of 16 stars across multiple sectors (32 light curves), cut into two-day windows resampled to 512 points, giving 324 real background windows; the train/test split is *by star* (12 for training, 4 held out) so no star appears in both.

Positive examples inject Mandel & Agol (2002) transit models (via `batman`) with randomised depth ($\delta \approx (R_p/R_\star)^2$), duration, impact parameter, and epoch; negatives are quiet windows and realistic artefacts. This yields 6,000 training and 4,000 test examples with injected signal-to-noise $\mathrm{SNR} \approx \delta\sqrt{n_{\mathrm{in}}}/\sigma$. A 1-D CNN maps each window to a transit probability (binary cross-entropy, Adam, 30 epochs, Apple-MPS); sensitivity is measured by ROC-AUC and by injection-recovery (completeness versus depth and SNR).

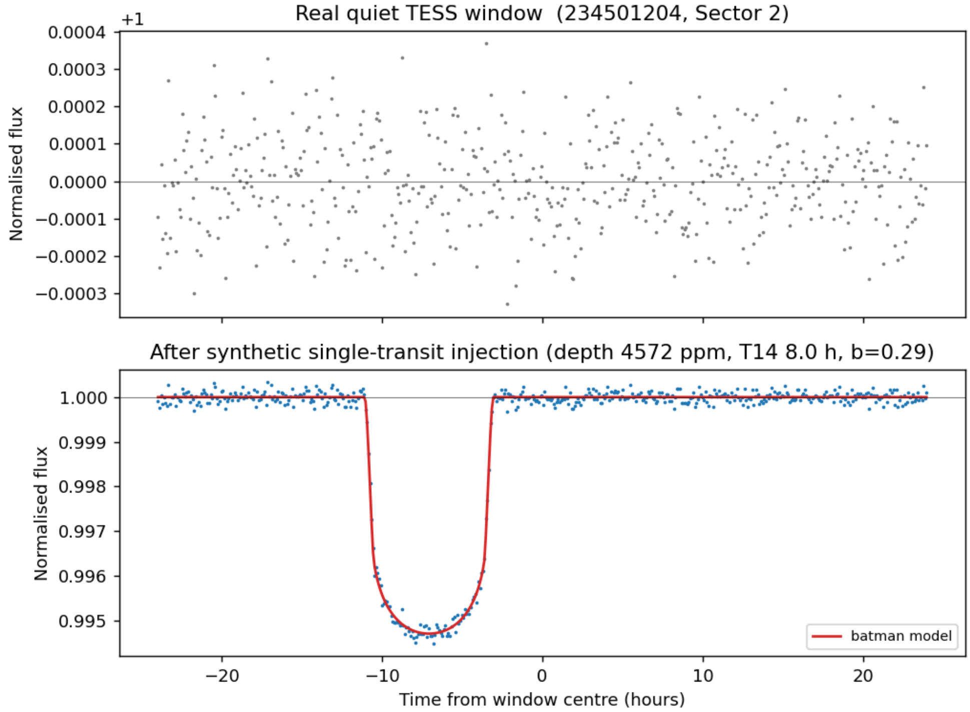

Figure 5.1: The training signal. Top: a real quiet TESS window (TIC 234501204, Sector 2). Bottom: the same window after injecting a synthetic single transit (depth 4572 ppm, duration 8 h, impact parameter 0.29); the `batman` analytic model (red) overlays the injected data (blue). Backgrounds are real; transits are synthetic and exactly labelled.

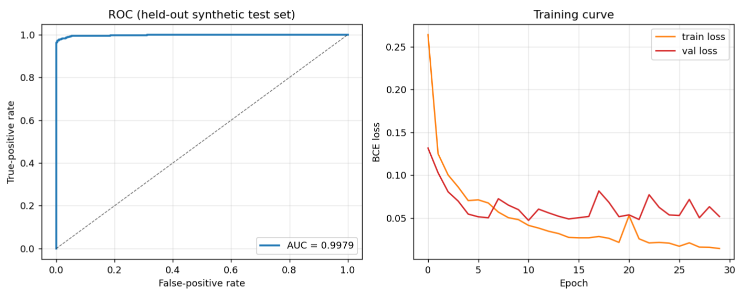

Figure 5.2: Receiver-operating-characteristic curve on the held-out, star-disjoint test set (AUC = 0.998); at the operating threshold the false-positive rate is 0.016 and the completeness is 0.977.

5.3 Caveats

The transits are synthetic; the method is validated by injection-recovery, and applying the trained detector to real data yields *candidates*, not detections. A single event has no period

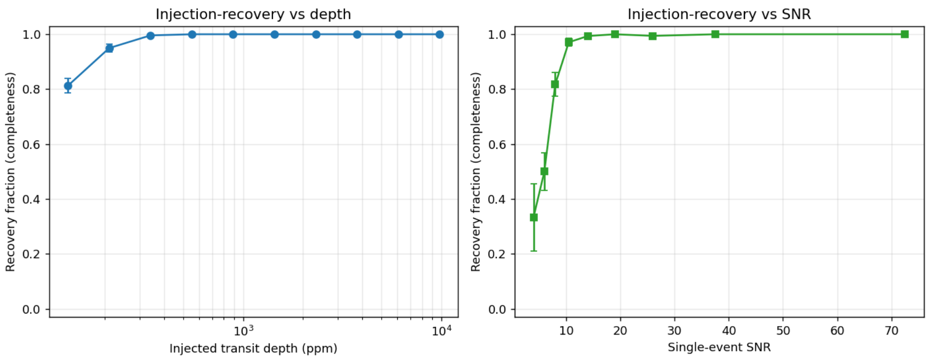

Figure 5.3: Injection-recovery. Completeness rises monotonically with signal strength (from ∼33% at SNR = 4, through ∼82% at SNR = 8, to essentially 100% above SNR ≈ 15) and with transit depth (∼81% at 131 ppm to ∼100% above 341 ppm) — the expected, physically sensible sensitivity of a transit search.

and no confirming second transit, so the real-data false-positive rate is high (grazing eclipsing binaries, stellar variability, systematics); strict vetting — odd/even and secondary-eclipse checks where possible, and neighbour-pixel analysis given TESS's ∼21 arcsec pixels — is essential. The 16-star background is a modest sample of real noise; a production search would draw backgrounds from thousands of stars.

References

1. G. R. Ricker et al. (2015). *The Transiting Exoplanet Survey Satellite (TESS).* JATIS 1, 014003. https://doi.org/10.1117/1.JATIS.1.1.014003

2. L. Kreidberg (2015). *batman: BAsic Transit Model cAlculatioN in Python.* PASP 127, 1161. https://doi.org/10.1086/683602

3. K. Mandel & E. Agol (2002). *Analytic Light Curves for Planetary Transit Searches.* ApJ 580, L171. https://doi.org/10.1086/345520

4. Lightkurve Collaboration (2018). *Lightkurve.* ascl:1812.013. https://docs.lightkurve.org

5. G. Kovács, S. Zucker, & T. Mazeh (2002). *A box-fitting algorithm in the search for periodic transits (BLS).* A&A 391, 369. https://doi.org/10.1051/0004-6361:20020802

6. N. L. Eisner et al. (2021). *Planet Hunters TESS I: results for the first two years.* MNRAS 501, 4669. https://doi.org/10.1093/mnras/staa3739