The Astrophysical Journal, 990:201 (23pp), 2025 September 10 https://doi.org/10.3847/1538-4357/adf2a1

The ATLAS Virtual Research Assistant

H. F. Stevance[1,2,3], K. W. Smith[1,2], S. J. Smartt[1,2], S. J. Roberts[4], N. Erasmus[5,6], D. R. Young[2,5], and A. Clocchiatti[7,8]
[1] Astrophysics Sub-department, Department of Physics, University of Oxford, Keble Rd., Oxford, OX1 3RH, UK; hfstevance@gmail.com
[2] Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, BT7 1NN, UK
[3] University of Sheffield, Astrophysics Research Cluster, Hicks Bldg., Broomhall, Sheffield S3 7RH, UK
[4] Department of Engineering Science, University of Oxford, UK
[5] South African Astronomical Observatory, Cape Town, 7925, South Africa
[6] Department of Physics, Stellenbosch University, Stellenbosch, 7602, South Africa
[7] Instituto de Astrofísica, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Macul, Santiago, Chile
[8] Millennium Institute of Astrophysics MAS, Nuncio Monseñor Sótero Sanz 100, Of. 104, Providencia, Santiago, Chile
Received 2025 April 11; revised 2025 July 17; accepted 2025 July 17; published 2025 September 10

Abstract

We present the Virtual Research Assistant (VRA) of the ATLAS sky survey, which performs preliminary eyeballing on our clean transient data stream. The VRA uses histogram-based gradient-boosted decision tree classifiers trained on real data to score incoming alerts on two axes: "Real" and "Galactic." The alerts are then ranked using a geometric distance such that the most "real" and "extragalactic" receive high scores; the scores are updated when new lightcurve data is obtained on subsequent visits. To assess the quality of the training we use the recall at rank K, which is more informative to our science goal than general metrics (e.g., accuracy, F1-scores). We also establish benchmarks for our metric based on the pre-VRA eyeballing strategy, to ensure our models provide notable improvements before being added to the ATLAS pipeline. Then, policies are defined on the ranked list to select the most promising alerts for humans to eyeball and to automatically remove bogus alerts. In production the VRA method has resulted in a reduction in eyeballing workload by 85% with a loss of follow-up opportunity $<0.08\%$. It also allows us to automatically trigger follow-up observations with the Lesedi telescope, paving the way toward automated methods that will be required in the era of LSST. Finally, this is a demonstration that feature-based methods remain extremely relevant in our field, being trainable on only a few thousand samples and highly interpretable; they also offer a direct way to inject expertise into models through feature engineering.

Unified Astronomy Thesaurus concepts: Sky surveys (1464); Transient detection (1957); Astrostatistics (1882); Interdisciplinary astronomy (804); Astroinformatics (78)

1. Introduction

The first two decades of the 21st century have seen a revolution in astronomers' ability to survey the sky on a large scale and in the time domain, with facilities such as Pan-STARRS (Panoramic Survey Telescope and Rapid Response System; N. Kaiser et al. 2002; K. C. Chambers et al. 2016), the Palomar Transient Factory (N. M. Law et al. 2009), ASAS-SN (All-sky Automated Survey for Supernovae; B. J. Shappee et al. 2014), the Zwicky Transient Facility (ZTF; E. C. Bellm et al. 2019), BlackGem (P. J. Groot et al. 2024), GOTO (Gravitational-wave Optical Transient Observer; D. Steeghs et al. 2022; M. J. Dyer et al. 2024), and ATLAS (Asteroid Terrestrial-impact Last Alert System; J. L. Tonry et al. 2018; K. W. Smith et al. 2020). These wide-field sky surveys have allowed astronomers to routinely find new transient events, which range from the common (a few thousand examples) thermonuclear (Type Ia) supernovae and core-collapse supernovae (CCSNe) (e.g., D. A. Perley et al. 2020; S. Srivstav et al. 2025, in preparation) to the rarer (a few hundred) tidal disruption events (TDEs; S. Gezari 2021) and superluminous supernovae (SLSNe; A. Gal-Yam 2019), as well as recently discovered optical counterparts of both gamma-ray bursts (S. B. Cenko et al. 2013) and fast X-ray transients (J. H. Gillanders et al. 2024). Until the advent of ZTF and ATLAS, counterparts to high-energy transients were typically found by focused follow-up but now that the whole sky can be scanned every 24–48 hr, optical afterglows are frequently found either without a high-energy trigger (D. A. Perley et al. 2025), or through post hoc association (e.g., B. Stalder et al. 2017).

Once transients are found in sky surveys, follow-up observations can be carried out and the science exploitation phase begins. These additional observations often require facilities with larger apertures and/or specialized instruments that are in high demand (e.g., Liverpool Telescope, I. A. Steele et al. 2004; the PESSTO and ePESSTO programs on the New Technology Telescope, S. J. Smartt et al. 2015; or X-shooter on the Very Large Telescope, J. Vernet et al. 2011). After overcoming the technical challenge of rapidly observing large areas of the sky with a rapid cadence (a few days), the field of transient astronomy transitioned from a target-limited regime to a resource-limited regime, where the number of transients far outweighed the availability of follow-up facilities. There began the new challenge of data curation and prioritization: how can we select the most promising or interesting alerts in a vast stream without overwhelming the science teams with data to manually eyeball? This usually starts with basic cuts (e.g., based on signal-to-noise), followed by cross-matching to astrophysical catalogs (D. R. Young 2023), real/bogus (RB) classification using convolutional neural networks (CNNs)

(e.g., T. L. Killestein et al. 2021; J. G. Weston et al. 2024), and finally eyeballing performed by humans to determine which alerts are indeed real and which need further attention. Even in surveys where this final step is handled with the help of citizen scientists (e.g., T. L. Killestein et al. 2024), they only contribute a small number of reported discoveries. In ATLAS, eyeballing requires between 200 and 400 alerts a day; a new step of automation was therefore required.

At this stage, early photometric transient classification—which uses as little lightcurve information as possible to infer likely spectroscopic classes—may seem like an attractive option. Some promising examples can be found in the work of D. Muthukrishna et al. (2019), who created RAPID, a recurrent neural network (RNN) aiming to classify 12 types of explosive transients as early as 2 days after the trigger, and in that of A. Gagliano et al. (2023), who presented a multimodal neural network using shallow learning on the image stamps and additional features to classify supernovae as early as 3 days after alert. Although preliminary classification can help in prioritizing transients for follow-up, these algorithms can only perform successfully in a stream that has been cleaned of other contaminants such as Galactic transients and leftover bogus detections, as they were trained on clean data sets (PLAsTiCC [The PLAsTiCC team et al. 2018] and the ZTF Bright Transient Survey data, respectively).

A different strategy is therefore required to clean the stream, one more suited to the task of triaging when little lightcurve information is known and many contaminants are present. A successful example of this is `BTSBot`, which flags potential bright extragalactic transients that are candidates for follow-up (N. Rehemtulla et al. 2024). Its classification is simpler than that in photometric transient classifiers (binary versus multi-class) but effective and adapted to the task of filtering data in a stream composed of many types of astrophysical events and leftover bogus.

In this paper we present a different approach to data curation in an impure transient stream. We call it the Virtual Research Assistant (VRA) because the strategy implemented in our design follows that of our human eyeballers. In Section 2 we summarize the design of the VRA, place it in the context of the ATLAS pipelines, and establish benchmarks we will use to assess the success of our algorithms. In Section 3 we present our data sets and the training of our models. In Section 4 we describe how the combined performance of our models and policies is evaluated before they are launched into production, and then report the in-production performance of the VRA. Further discussion can be found in Section 5 and we conclude in Section 6. In addition to this paper, we point the reader to the manual (H. Stevance 2025a) for further details on earlier prototypes of the VRA and for up-to-date information on the current VRA version and eyeballing policies. All the data and code used for the training and analysis presented here can be found in the VRA v1 code and data release on Zenodo: doi:10.5281/zenodo.15195392 (H. Stevance 2025b).

2. Overview

2.1. ATLAS Transient Searches

The ATLAS sky survey (J. L. Tonry et al. 2018) is composed of four 0.5 m telescopes: two in Hawaii, one in Chile, and one in South Africa. Two filters are used for observations: the cyan (c) filter (420–650 nm) used during dark time and the orange (o) filter (560–820 nm) used during bright time. In survey mode ATLAS performs four 30 s exposures of tile or sky pointing, each separated by roughly 15 minutes, a strategy motivated by the main science case of ATLAS (the discovery and follow-up of near-Earth objects; A. N. Heinze et al. 2021). The individual frames are detrended and calibrated (astrometrically and photometrically) on-site for each unit and the data are transferred to Hawaii. Difference imaging with respect to the ATLAS wallpaper is carried out, after which all sources with significance greater than 5σ are cataloged. These detection catalogs and the reduced and calibrated images are transferred to transient servers in Queen's University Belfast (K. W. Smith et al. 2020). The catalog files of the difference image detections contain about $\mathcal{O}(10^7)$ sources, all of which are ingested into a relational database. Before the development of the VRA, the transient alert processing (as summarized in K. W. Smith et al. 2020) was as follows:

1. Quality cuts: three or more good-quality, cospatial detections at significance of 5σ or greater within 1 night are required to define an object.
2. Astrophysical cross-matching with SHERLOCK (D. R. Young 2023). Known variable stars are removed from the stream and contextual information is added, such as potential host cross-matching, angular distance to potential hosts, and redshift (if known) (see Sherlock documentation for further details).
3. RB classification using a CNN (J. G. Weston et al. 2024). The 20×20 central pixels of each difference image receive a score between 0 (bogus) and 1 (real). Below a 0.2 threshold the alerts are not sent for human eyeballing—they are directly labeled as garbage (but not deleted from the database).
4. Humans eyeball the data to classify them into four broad categories: "good" (extragalactic transient), "attic"/"Galactic," "garbage," and "proper motion" (PM; alerts due to stars moving between the time of observation and the date at which the wallpapers were constructed). To do this, humans have access to a broad range of (multimodal) data: stamps of the wallpaper, observation, and difference imaging; lightcurve; and contextual information.

We can gauge the workload of the eyeballers using the first data set gathered for VRA training between 2024 March 27 and 2024 August 13. Over that period a total of 40,802 objects were presented to humans for scanning, averaging nearly 300 objects per day. As we can see in Figure 1, nearly 90% of those objects were labeled as garbage or PM, the rest being nearly evenly split between the attic (Galactic) and good (extragalactic) categories.

Figure 1 encapsulates the problem we will address in this paper: the real, extragalactic transient sources are still only a few percent (5.5% over this period) of the objects that a human on duty will scan through manually. Further automation of the eyeballing process is challenging, at least in part because human eyeballers are able to triage the objects with very little lightcurve information. As we can see in Figure 2, a large portion of the objects (90%) is labeled in less than 48 hr. Hence human scanning is quick and effective but the ratio of good to reject objects makes it an inefficient use of scientist time.

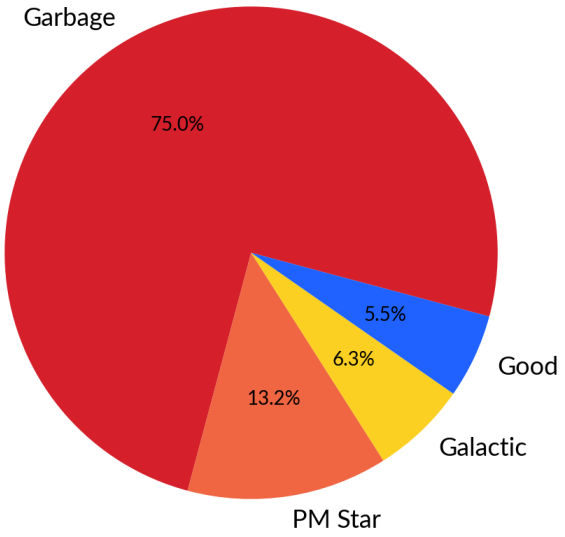

Figure 1. Alert type distribution in the ATLAS eyeball list between 2024 March 27 and 2024 August 13 for a total of 40,802 alerts, all eyeballed by humans and predating the introduction of the first VRA prototype in production. See the label description in Section 3.1.

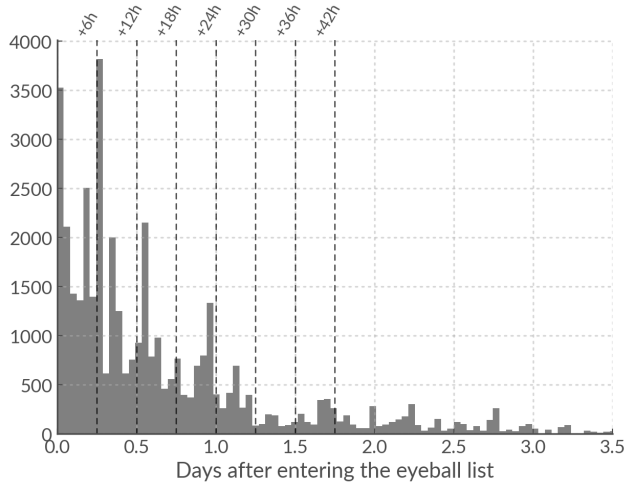

Figure 2. Histogram of the time delay between an alert entering the eyeball list and a human making a classification between 2024 March 27 and 2024 August 13.

2.2. Scope and Benchmark

The primary goal of this project is to minimize eyeballer workload without compromising extragalactic transient identification. Henceforth, we shall define the objects that make it through the basic filtering as "alerts" for which we aim to make the process more efficient. A "simple" way (conceptually, albeit not necessarily technically) to do this is to order the eyeball list with alerts that are most likely to be "good" at the top. If we can provide a ranking that is robust enough to guarantee complete recovery of the "good" alerts in the top X % of the list (X to be determined after training), we can then crop the bottom of the eyeball list. In essence, that is the strategy implemented with the RB score, where all alerts below 0.2 are automatically labeled as garbage. In practice, however, the RB score ordering is not very effective beyond the initial crop.

As can be seen in Figure 3 (left panel), the distribution of the RB scores for the garbage alerts has a secondary peak at an RB score of 1 (the primary peak at 0 is not shown as the plots only show scores for eyeballed alerts with RB score > 0.2), which leads to confusion in the high RB score regions. In the right panel of Figure 3 we also show the recall at rank K (R@K) obtained from using the RB score to rank the eyeball list. The R@K is defined as

$$\mathrm{R@K} = \frac{N \text{ relevant alerts in top } K}{N \text{ relevant alerts}}. \quad (1)$$

Since our scientific focus is on extragalactic transients ("good" list) these are the objects considered relevant for R@K calculations. Our goal is to create models that result in an R@K curve that is steeper than that in Figure 3 and ideally reaches 100% recall closer to the top of the list (currently only beyond a fraction of 0.8, or 80% down the list). Inspired by the area under the receiver operating characteristic (AUROC) metric, which is classically used when training machine learning (ML) models, we define the area under the R@K (AuRaK) as one of our model evaluation metrics. For the full data set ordered by RB score, we obtain an AuRaK = 0.88. Any model we create should exceed this value and show a steeper rise; otherwise it would provide no improvements compared to the current strategy.

The other issues with ordering and selection by RB score are that a simple, binary RB score does not discriminate between Galactic and extragalactic transients, and it does not capture the new information provided by new lightcurve points obtained in subsequent observations (whether it be a detection or a nondetection). We need to create a system that can update the ranking of an alert when new data is gathered.

2.3. A Transient-agnostic Score Space

One of the first steps in the design of the VRA eyeballing system was to perform interviews with the most experienced members of the eyeballing team to ask what questions they ask themselves prior to making decisions and what pieces of data they use to answer them. At this stage of eyeballing, there are only three important questions to be answered:

1. Is it real?
2. Is it Galactic?
3. Is it fast?

The transient classification, even a broad version of it (e.g., is it a Type II versus a Type Ia), is not a major consideration in the first few days of an alert because in most cases the data is simply insufficient to make a reliable, informed decision. Therefore a transient-specific classifier is not adapted to the task at hand.

Additionally, the three questions highlighted above conveniently define a score space within which all varieties of transients and boguses live (see Figure 4). Since it is not specific to one class of transients we call it "transient-agnostic."[9] Overall the VRA is designed to follow a similar strategy to the human eyeballers: we create scoring algorithms that take data from the stream and provide a real score (p_{real}) and a Galactic score (p_{gal}). In the end the idea of assigning a

[9] However, we note that our implementation in the ATLAS VRA, because it will be assessed using metrics that favor extragalactic transients, will be biased toward performing well on extragalactic transients (see Sections 3 and 4.1).

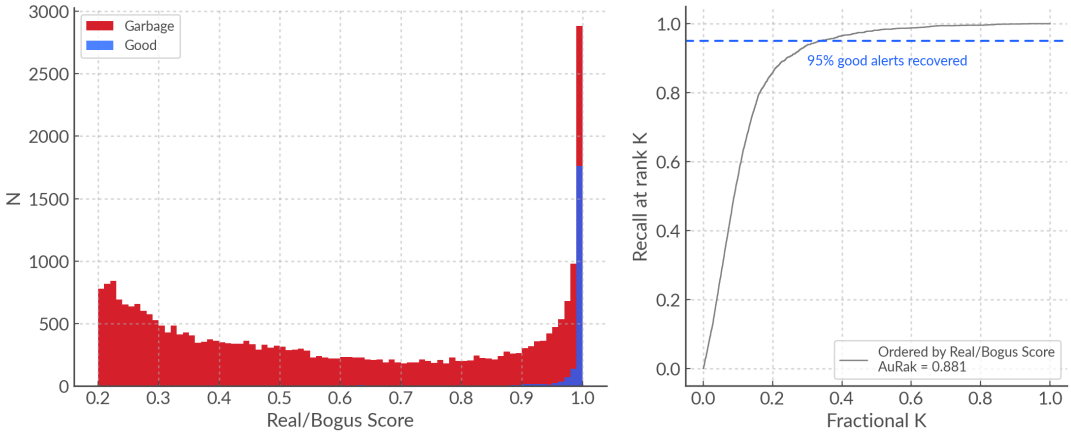

Figure 3. Left: distribution of the RB score for the human-labeled "garbage" and "good" alerts over the period 2024 March 27–August 16. Right: R@K for the data set ordered by RB score. When ordering by RB score the eyeballers would have to, on average, eyeball the top 35% of the list to recover 95% of the good objects (amounting to 5.3% of the list). To recover 99% of the good objects, 60% of the list (ordered by RB score) must be eyeballed.

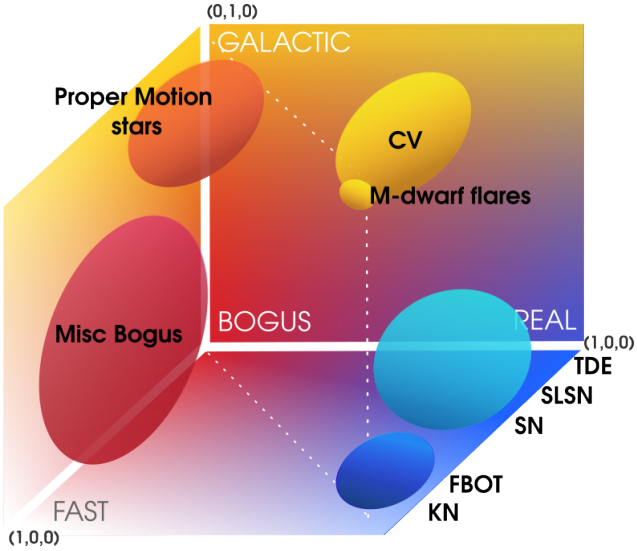

Figure 4. A toy representation of our transient-agnostic score space, defined by three dimensions: real (x), Galactic (y), and fast (z). The latter relates to the timescale on which the transient lightcurve evolves—it is not used in this iteration of the VRA and is a more subjective quantity (see discussion in Section 5). Nonetheless we can conceive a 3D space where all types of transients and bogus alerts can be found. Cataclysmic variables for example are fast (they can rise by several magnitudes within a day) and Galactic transients; kilonovae and fast blue optical transients are also rapidly evolving but extragalactic in nature. TDEs and SLSNe are extragalactic too but slower evolving than supernovae. As for the bogus alerts, artifacts from high-PM stars are Galactic in nature while other miscellaneous bogus events (e.g., from trailing) may show a variety of behavior. In the specific case of ATLAS (see Appendix A) most of our miscellaneous bogus alerts are highly correlated with the Galactic plane—this may not be a behavior that extends to other surveys.

fast score ($p_{\rm fast}$) was deprioritized for this version of the VRA since performance was found to be satisfactory (see Section 4.1), but future iterations of the ATLAS VRA (or other bots in future surveys) may want to use all three axes.

Once alerts have been placed in the score space, we must then rank them from most relevant to least relevant in the eyeball list. The advantage of a transient-agnostic score space is that ordering the alerts by relevance can be adjusted for different science cases without having to modify (retrain) the scoring algorithms. Those looking for Galactic alert candidates would want alerts near the ($p_{\rm real} = 1$, $p_{\rm gal} = 1$, $p_{\rm fast} = 1$) coordinates to be ranked the highest, while supernova astronomers would favor alerts in the regions corresponding to ($p_{\rm real} = 1$, $p_{\rm gal} = 0$).

The details of how the scoring algorithms are trained can be found in Section 3.

2.4. Eyeballing Policies

Using the real and Galactic scores we calculate two properties: the VRA score,[10] which ranges from 0 to 10 and is used to rank alerts from least to most real/extragalactic, and the Galactic flag, a Boolean (true or false) that identifies alerts as being likely to be Galactic.

The VRA score measures the distance to the (1, 0) coordinates in score space, normalizes it, makes shorter distances yield a high score, and multiplies the result by 10 to obtain a score between 0 and 10. It is calculated as follows:

$$\mathrm{VRA}_{\rm score} = 10 \times \frac{\sqrt{f^2+1} - \sqrt{(1-p_{\rm real})^2 + (f \times p_{\rm gal})^2}}{\sqrt{f^2+1}} \quad (2)$$

where f is a scaler applied to the Galactic axis to better separate the "garbage" and "PM" distributions from real events (see for example Section 3.3.2). The scaler f is a parameter that allows us to tune the separation between distributions in score space. In the current version of the VRA $f = 0.5$, which allows greater separability between the bogus classes ("garbage," "PM") and the real classes ("Galactic," "good")—this is most easily seen in Section 3.3.2. The intuition behind this choice is as follows: for our science goals, we are more tolerant of confusion between Galactic and extragalactic transients than of confusion between real alerts (of any kind) and bogus alerts. For a discussion on the choice of f see Section 5. If an alert has a cross-match to the Transient Name Server (TNS; A. Gal-Yam 2021) we automatically upgrade its rank to 10. The eyeballers are tasked with inspecting objects with scores > 7.

In addition, we calculate a Galactic flag that measures the distance to the (1, 1) coordinates of score space (with a scalar $f = 0.9$) and returns "true" if that distance is <0.4. Objects that

[10] Note it is often called the rank in our internal codes and databases.

do not meet the VRA score threshold of 7 but do get flagged as potentially Galactic are moved to a separate "Galactic candidate" list to be eyeballed with lower priority. Finally, for objects with $D < 100$ Mpc, the VRA scores are used to flag high-rank objects for automated follow-up (see Section 6) but eyeballers are still tasked with looking at all incoming objects. At the end of each ingest cycle a slackbot is triggered and presents the eyeballers with three tiers of eyeballing priorities: fast-track (immediate), extragalactic (within the hour), and Galactic (within 24–48 hr).

2.5. Auto-garbaging Policies

Using the VRA score calculated with Equation (2) we select alerts for auto-garbaging if they meet the following criteria (at the time of writing): on day 1 their rank $\mathrm{VRA}_{\mathrm{score}} < 1$; on the second visit the maximum $\mathrm{VRA}_{\mathrm{score}} < 2$; and on the third visit and beyond the mean $\mathrm{VRA}_{\mathrm{score}} < 3$.

The general form of these policies was chosen when the VRA was first added to production in 2024 August, and further evaluation of the policies such as described in Section 4.1 led to the specific values presented here. It is worth reemphasizing that "garbage" is a label in the ATLAS Transient Server Database and a list to which the alerts are moved. Data are not deleted from the database.

Finally, note that all policies are subject to change over time as new versions of the VRA may be trained or policies revisited. For up-to-date information regarding the VRA policies and version please see the most up-to-date version of the technical manual (H. Stevance 2025a) or the VRA website.[11]

2.6. Monitoring

The operations of the VRA are monitored weekly. Every Friday a report is sent to a Slack channel summarizing the following information: the number of objects that entered the eyeball list (RB score > 0.2); the number of objects that were eyeballed by humans; the number of objects reported to TNS; a pie chart showing the distribution of labels for the past week; and the number of potential VRA misses.

Potential misses are defined as alerts that would have not met the VRA rank threshold but whose rank was raised to 10 by a cross-match to TNS. To further monitor potential misses we also have deployed a bot that cross-matches the "garbage" list items of the past week to TNS. TNS items can still land in the "garbage" list if their RB score was lower than 0.2 or if they met the VRA auto-garbaging policies before they were reported to TNS.

Finally, a purgatory sentinel runs every day to flag any alerts that have not been eyeballed or auto-garbaged but are more than 15 days old because the day N models are only trained with data up to day 15 (see Section 3.3) and we do not trust scores predicted out of distribution. We have not found this workload to be substantial (only a handful of objects) so we have not found the need to add additional eyeballing or garbaging policies. These are then eyeballed by a human to make a final decision.

These bots send Slack alerts and record their reports onto CSV files, which can be inspected at a later date. These regular checks have been crucial to development and will allow us to monitor the VRA for a decrease in performance in the future, which could occur as a result of data drift and could call for a retraining of the scoring algorithms.

[11] https://heloises.github.io/atlasvras/index.html

3. Real/Galactic Scoring

Placing alerts in our score space defined in Figure 4 is done by using two binary classifiers: an RB and a Galactic/extragalactic classifier. Additionally, we differentiate between alerts that are newly added to the eyeball list (day 1) and those that are receiving additional lightcurve information on subsequent days by creating day 1 and day N models. This is motivated by the fact that, although a majority of our alerts are classified on day 1 (Figure 2), this is partially skewed because many alerts are obviously bogus and do not require further data to make a decision. When establishing whether an alert is extragalactic or Galactic, waiting for additional lightcurve data is not uncommon in eyeballing, and the VRA must be able to use this new information. The day N models are trained on data ranging from day 2 to day 15 and include additional features (see Table 1 and Section 3.2) to capture informative lightcurve evolution that will help our real and Galactic classifiers. Overall we have four binary classifiers: real, day 1; Galactic, day 1; real, day N; and Galactic, day N.

Our classifiers are trained using data taken from the stream and labeled by our eyeballers. The labels and notable caveats are described in Section 3.1; the data set and features are described in detail in Section 3.2; and then the training of our models is presented in Section 3.3.

3.1. The Labels

There are four alert classification categories, which we use as follows:

1. *"Garbage."* This is a broad category that encompasses most types of bogus alerts such as trails, star spikes, bad point-spread function, detector issues, bad subtractions in crowded fields, and those associated with bright galaxy cores. This category is used to provide samples with label $p_{\mathrm{real}} = 0$.
2. *"PM."* This is used to separate bad subtractions that are specifically suspected to have occurred as a result of the drift of a star compared to its position in the ATLAS wallpaper. This category provides samples with labels $p_{\mathrm{real}} = 0$, $p_{\mathrm{gal}} = 1$.
3. *"Attic/Galactic."* The attic is a list in the ATLAS transient server (see K. W. Smith et al. 2020) used to store real alerts that do not belong in our "good" list. This contains mostly Galactic events (cataclysmic variables (CVs), stellar flares, and stellar variability) and we use this category to create training samples with labels $p_{\mathrm{real}} = 1$, $p_{\mathrm{gal}} = 1$.
4. *"Good."* This list is dedicated to extragalactic transient alerts (supernovae, TDEs, SLSNe, and fast blue optical transients (FBOTs), but not active galactic nuclei (AGNs)) and the samples drawn from this are given labels $p_{\mathrm{real}} = 1$, $p_{\mathrm{gal}} = 0$.

As with any real sample, the labeling is not pure and there are known areas of confusion or mislabeling. The "PM" category is a more recent addition to the web server and although it predates the start of our data gathering for the VRA project some eyeballers would place examples of "PM" stars

Table 1
Features Used by the Day 1 and Day *N* Models

Model	Column Name	Type	Description
day 1 + *N*	Nnondet_std	float	Standard deviation of the number of nondetections between detections
day 1 + *N*	Nnondet_mean	float	Mean of the number of nondetections between detections
day 1 + *N*	magdet_std	float	Standard deviation of the magnitude of the historical detections
day 1 + *N*	DET_Nsince_min5d	float	Number of detections between phase −5 days and day 1
day 1 + *N*	NON_Nsince_min5d	float	Number of nondetections between phase −5 days and day 1
day 1 + *N*	DET_mag_median_min5d	float	Median magnitude of the detections between phase −5 days and day 1
day 1 + *N*	log10_std_ra_min5d	float	log 10 of the standard deviation of the R.A. in the detections from phase −5 days
day 1 + *N*	log10_std_dec_min5d	float	log 10 of the standard deviation of the decl. in the detections from phase −5 days
day 1 + *N*	ra	float	R.A.
day 1 + *N*	dec	float	decl.
day 1 + *N*	rb_pix	float	RB score from the CNN (J. G. Weston et al. 2024)
day 1 + *N*	z	float	Spectroscopic redshift (if known, else NaN)
day 1 + *N*	photoz	float	Photometric redshift (if known, else NaN)
day 1 + *N*	ebv_sfd	float	Extinction $E(B-V)$ calculated using dustmaps SFD
day 1 + *N*	log10_sep_arcsec	float	log 10 of the projected separation between the best-matched source (in arcseconds)
day 1 + *N*	SN	bool	[PRUNED] (Sherlock classification) If SUPERNOVA
day 1 + *N*	NT	bool	[PRUNED] (Sherlock classification) If NUCLEAR TRANSIENT
day 1 + *N*	ORPHAN	bool	[PRUNED] (Sherlock classification) If ORPHAN
day 1 + *N*	CV	bool	(Sherlock classification) If CATACLYSMIC VARIABLE
day 1 + *N*	UNCLEAR	bool	[PRUNED] (Sherlock classification) If UNCLEAR
day *N* only	max_mag	float	Maximum (median) magnitude seen since phase −5 days
day *N* only	max_mag_day	float	Day of the maximum magnitude
day *N* only	DET_N_total	float	Number of detections since phase −5 days
day *N* only	NON_N_total	float	Number of nondetections since phase −5 days
day *N* only	DET_mag_median	float	Median magnitude of the detections since phase −5 days
day *N* only	NON_mag_median	float	Median magnitude of the nondetections since phase −5 days
day *N* only	DET_N_today	float	[PRUNED] Number of detections seen today
day *N* only	NON_N_today	float	[PRUNED] Number of nondetections seen today

Note. The column names are those used in the code and data release.

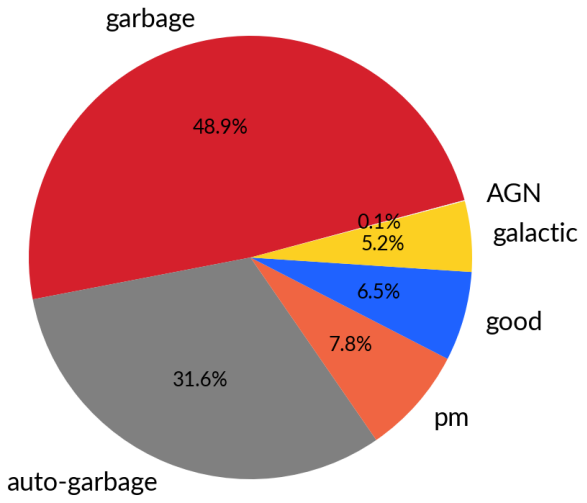

Figure 5. Alert type distribution in our full data sets spanning 2024 March 27 to 2025 January 22. Some of these alerts were re-eyeballed during development as their human labels were discrepant with their location in score space. A few AGNs were found in the "Galactic" ("attic") alerts and marked as such.

in the garbage. Another area of confusion arises in the "attic," which contains duplicate good objects and some AGNs in low numbers. Re-eyeballing of the data during development allowed us to find some of these alerts and remove their $p_{\rm gal} = 1$ labels (some contamination may remain; see H. Stevance 2025a for details).

3.2. Data and Features

The data used to train the models presented in this paper were gathered from the eyeball list between 2024 March 27 and 2025 January 22. These do not reflect the full extent of the bogus properties in the ATLAS stream and are solely intended to train a model that works downstream of previous automation steps (points 1–3 in Section 2.1).

The VRA underwent several rounds of prototyping and the data was divided into sub–data sets. The first sub–data set was gathered between 2024 March 27 and 2024 August 13, during which no VRA prototypes were actively participating in eyeballing. These data best reflect the eyeballer workload and decision speed, although we note that they are limited in time to a period of four and a half months during which the Galactic center is very visible to the Chilean and South African ATLAS units. The second sub–data set was gathered between 2024 August 18 and 2025 January 22. This is the first data set that is impacted by the VRA, which is reflected in the large fraction of data that are labeled as "auto-garbage."

The full data set used here contains 75,129 alerts with the label distribution shown in Figure 5. The raw JSON data and cleaned data frames are available alongside the codes used to clean the data set and make the features described in the following sections (H. Stevance 2025b). The features used by

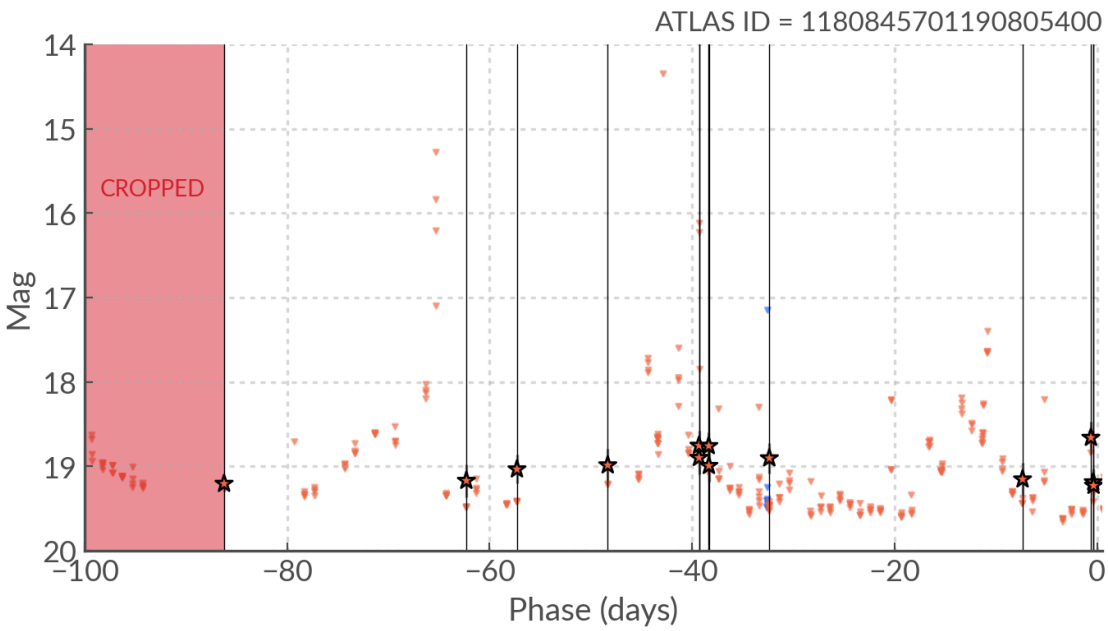

Figure 6. Visualization of the lightcurve history used to create the VRA features. The star markers show detections (any filter) and the triangles show the 5σ limiting magnitude in the ATLAS frame of each nondetection. The time axis is given with respect to the time of first alert ($t = 0$). Each detection is also highlighted with a vertical line to show how historical detections segment the lightcurve—it is within these segments that we count the nondetections to then calculate the mean and standard deviation for the `Nnondet_std` and `Nnondet_mean` features. Also in red we show that every measurement before the first detection in our −100 day window is cropped (red overlay).

the day 1 and day N models are summarized in Table 1, and the feature distributions are shown in Appendix A.

3.2.1. Long-term Lightcurve History (−100 days)

One of the characteristics that eyeballers look for in the alert lightcurves is historical detections, as they can indicate recurrent activity or outbursts (real) or regular bad subtractions or artifacts at that location (bogus).

To attempt to capture these behaviors we calculate three features: the mean and standard deviation of the number of nondetections between each historical detection (`Nnondet_mean`, `Nnondet_std`), and the standard deviation of the magnitude of these historical detections (`magdet_std`). Here a historical detection is defined as any detection that occurred within −100 days of entering the eyeball list. This includes lone detections that would not pass the quality cuts that require a minimum of three detections within a single night. To calculate the features we first crop every data point before the first historical detections, within our chosen time window of −100 days, in order to anchor our count of nondetections. To illustrate this process we show in Figure 6 an example of a garbage alert with spurious detections. The mean and standard deviation of the number of nondetections between each historical detection are 19.1 and 27.4.

For feature calculations we ignore the filter information, which means that orange and cyan magnitudes are considered together when calculating, e.g., the standard deviation. Also it is worth stating that magnitudes being a logarithmic transformation of the flux, taking the standard deviation of a series of magnitudes is not the correct way to formally assess their variability. Nonetheless we use such features as they are informative and fast to compute, but we emphasize that they are not strictly physical measurements and they should not be used outside of this context—certainly not if deriving astrophysical quantities.

3.2.2. Recent Lightcurve History (−5 days)

The recent lightcurve history is defined as the data captured within −5 days of the alert entering the eyeball list. This is of particular interest as real transients that were rising but faint in preceding days may have shown one or two detections but would have fallen short of our requirement for alerts to have three out of four 5σ detections in one night to enter the data stream. We choose 5 days before alert as a cutoff since given the ATLAS cadence it usually corresponds to an additional one to two previous visits, which from eyeballing experience are where lower-signal detections begin to be visible for rising extragalactic transients.

We record three recent lightcurve history features: the number of detections (`DET_Nsince_min5d`), the number of nondetections (`NON_Nsince_min5d`) seen in the 5 days preceding and including the first alert, and the median magnitude of the detections over that period (`DET_mag_median_min5d`).

3.2.3. New Lightcurve Information for the Day N Models

For the day N models, we calculate four additional lightcurve features. These are designed to try and capture the new lightcurve information gathered by new telescope visits for sources that were ambiguous and not classified by eyeballers on day 1. The four additional features are the total number of detections (`DET_N_total`) and nondetections (`NON_N_total`) since −5 days, the minimum (maximum brightness) magnitude value recorded so far (`max_mag`), and the day (phase) on which that magnitude value was recorded (`max_mag_day`). Again here all filters are considered together and there is no distinction between a cyan maximum

and an orange maximum. We do not record which filter the maximum was recorded in as it would not be informative for VRA classification since the orange and cyan filters alternate based on the phase of the Moon rather than on characteristics related to the transients themselves.

We also do not attempt the fit the lightcurves to find the peak magnitude. Previous work has shown that the simpler feature of the minimum magnitude and the day on which it occurred is surprisingly informative (N. Rehemtulla et al. 2024) and we also find that to be the case (see Sections 3.3.2 and 3.3.3).

We only calculate these features until a maximum phase of 15 days after the initial alert.

3.2.4. Sky Location and Extinction

Some of the most important contextual features used are the on-sky positions (R.A. and decl.) and the extinction ($E(B-V)$; `ebv_sfd`) (see Sections 3.3.2 and 3.3.3). This is because they define whether the source is associated with a crowded area of the sky (Galactic plane) that is prone to bad subtractions (see Figure 19 in Appendix A). High values of extinction reduce the likelihood of an alert being an extragalactic source. In the extreme, the highest values of foreground extinction (e.g., $E(B-V) \gtrsim 1$) preclude extragalactic sources simply due to the limit they place on the absolute magnitude.

3.2.5. Scatter on the Sky

Another important piece of information leveraged by the eyeballers is the scatter in the R.A. (`log10_std_ra_min5d`) and decl. (`log10_std_dec_min5d`) of the individual exposures. Typically a well-localized alert, with little scatter in its position centroid, will be associated with a real alert. The converse is not necessarily true, however, as real transients can and do occur in images that show jitter or trailing. We apply a log 10 transformation to these features to obtain a more symmetrical distribution and since these values are never 0 there is no risk of getting undefined values.

3.2.6. SHERLOCK Cross-matching

Finally we also record features relating to cross-matching with astrophysical catalogs. This cross-matching is already performed upstream with SHERLOCK (D. R. Young 2023), which is a contextual classifier that uses boosted decision trees to perform rapid cross-matching to existing astrophysical catalogs for transient surveys used in ATLAS and the Lasair data broker (R. D. Williams et al. 2024). It adds features to the data stream related to host/source cross-matching (such as angular distance and redshift), as well as a classification based on the features in the cross-match.

For our purposes we use four of these precalculated features: the logged separation in arcseconds to the cross-matched source (`log10_sep_arcsec`), the spectroscopic redshift (`z`), the photometric redshift (`photoz`), and finally the CV flag. This flag indicates that a known classified source is within 0.″5 of our transient candidate.

In this final version of the VRA we do not use the SN (suspected supernova), ORPHAN (no host), NT (nuclear transient), or UNCLEAR SHERLOCK flags because they induce confusion in the models (see Sections 3.3.2 and 3.3.3). That is due in part to the fact that many Galactic stars are tagged as extended sources in some catalogs, leading both SHERLOCK and our models to confuse Galactic transients with potential supernovae.

3.3. The Models

3.3.1. Histogram-based Gradient-boosted Classifiers

To place our alerts in score space we created two classifiers: an RB classifier and a Galactic/extragalactic classifier (four if we include their day 1 and day N variants). We chose to use a gradient boosting method (J. H. Friedman 2001) called histogram-based gradient-boosted decision trees ((H)GBDT).

Lightcurve data in its raw state is a time series, which is not handled well by such feature-based classifiers. This is why we devised an array of features to capture long- and short-term lightcurve behavior. Although there exist other forms of ML methods such as neural networks (especially RNNs) that could use the raw lightcurve as input, there are two reasons why we did not choose to use such models. First, neural networks are data-hungry. To be trained effectively they require data sets of order 50× the number of parameters in the model (A. Alwosheel et al. 2018). Although we appreciate that this is somewhat a "rule of thumb," other studies indicate that the GBDT family can provide strong baselines across a wide range of data set sizes (D. McElfresh et al. 2023). The fact that (H)GBDT models can perform well with only a few thousand examples allows us to proceed without the need for data augmentation to artificially increase the number of data samples, as is often the case when using neural networks with large numbers of parameters.

Second, the lightcurve information is not rich on the timescales considered here, and it is difficult to represent astrophysical time series in classical ML and statistical tools. For example they are not built to handle nondetection information (a nondetection is not the same as no observation). On the whole we prefer to engineer our own features based on our understanding of the lightcurve information and our goals.

Finally, the use of feature-based methods allows for easier interpretability of the models, which makes diagnosing potential issues easier (see later our discussion of AT 2024lwd in Section 4, Section 4.1.1, and Figure 25 in Appendix A).

We use the `scikit-learn` (F. Pedregosa et al. 2011) implementation of (H)GBDT, which is based on LightGBM (G. Ke et al. 2017). There are several advantages to the histogram-based approach; of major importance to us is the native handling of null values. In most ML models null values need to be imputed; in (H)GBDT the feature values are discretized into histograms with (typically) 256 bins, 255 of which are used for numerical values and the final bin for null values.[12] Another advantage (and the main reason this improvement on the original algorithm was devised) is the speed. This only becomes a concern when dealing with tens of thousands of training samples, which is barely the case here but could become important in the future.

3.3.2. Day 1 Models

We first split the data into a training set and a validation set, at ratios of 0.85 and 0.15 of the full data set. The training set was then balanced by subsampling overly represented classes. As mentioned above, the data set considered here is a combination of two sub–data sets, the training/validation split

[12] See manual page for (H)GBDT in scikit-learn.

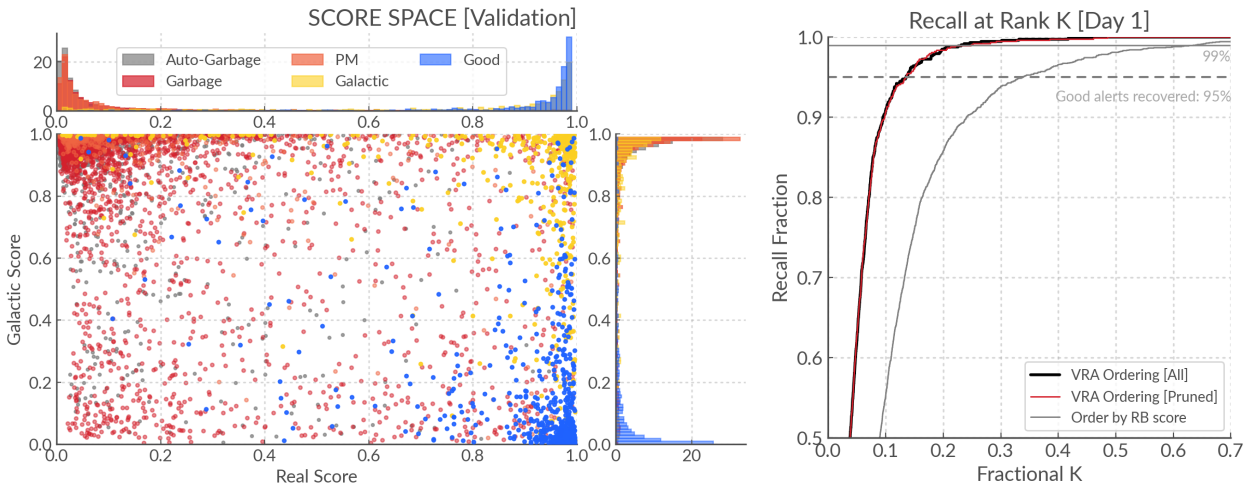

Figure 7. Score space and R@K of our validation data set for the day 1 models, color-coded according to their human-given labels. The score space is that obtained with the models trained excluding the pruned features (see Table 1). The R@K plot shows both the R@K curves obtained for the models trained on all the features and those trained excluding the pruned features. For comparison with our benchmark in Figure 3 we show the R@K curve obtained when ordering by the RB scores calculated by the CNN.

Table 2
Day 1 Model Training and Validation Set Label Distribution

Label	Training	Val.
Good [$p_{\rm real} = 1$; $p_{\rm gal} = 0$]	4249	745
Galactic [$p_{\rm real} = 1$; $p_{\rm gal} = 1$]	2899	457
PM [$p_{\rm real} = 0$; $p_{\rm gal} = 1$]	2736	905
Garbage [$p_{\rm real} = {\rm NaN}$; $p_{\rm gal} = 0$]	2921	5380
Auto-garbage [$p_{\rm real} = {\rm NaN}$; $p_{\rm gal} = 0$]	1600	3571
TOTAL	14,405	11,058

and resampling of which were performed individually. Additionally, the resampling was performed before parts of the data sets were re-eyeballed. The full historical details of data gathering and re-eyeballing can be found in the online data release (H. Stevance 2025b) and the technical manual (H. Stevance 2025a). Overall the training set is not fully balanced, as we can see in Table 2, but it is not largely dominated by bogus alerts as are the full data set and the validation set.

Both the $p_{\rm real}$ and $p_{\rm gal}$ scorers were trained with an `l2_regularization` parameter of 10, a `class_weight` parameter set to "balanced," and a `random_seed` = 42. The $p_{\rm real}$ scorer was trained with a `learning_rate` of 0.1 (default) while the $p_{\rm gal}$ scorer `learning_rate` was set to 0.2. In an earlier prototype, hyperparameter optimization using a grid search was conducted on the `learning_rate` and `l2_regularization` values. We found very little difference in performance between most (reasonable) values of these parameters. The performance gains were too marginal to justify the time and computational cost of rerunning the hyperparameter search and follow-up tests on the AuRaK and eyeballing policies and subsequent trainings of the VRA. The details of our hyperparameter searches during development can be found in H. Stevance (2025a) and the code is available in the data release (H. Stevance 2025b).

In Figure 7 we show the score space and R@K curve obtained with our day 1 models. We can see good separation of the real classes ("good" and "Galactic") from the bogus classes ("PM" and "garbage"), and the extragalactic transient ("good") distribution is well confined to the bottom right corner of the score space. The "Galactic" samples are also well separated and found in the top right corner of the plot where we expect. The garbage samples naturally concentrate toward the top of the score space, which is unsurprising as "garbage" alerts are more likely to occur in crowded fields (as can be seen in Figure 19 in Appendix A, they partially track the Galactic plane).

We can also compare our day 1 models' performance in ordering the alerts using Equation (2) to our benchmark defined in Section 2.2 (right panel of Figure 7). We get 95% (99%) recall of the "good" objects within the top 15% (25%) of the list. To achieve the same recovery of "good" alerts when ordering with the RB score from the CNN RB classifier, we would have to scan 35% (>60%) of the list. The AuRaK has also increased from 0.88 for our benchmark to 0.951 with our day 1 models.

Then, we can quantify how informative each of our features is using a technique called permutation importance, which consists in scrambling one feature at a time and retraining a model to evaluate the decrease in prediction accuracy. In the present case we performed 10 iterations (for each feature) using the `scikit-learn` implementation; the results for both our real and Galactic scoring models are presented in Figure 8. We can see that the SHERLOCK flags SN, UNCLEAR, ORPHAN, and NT do not result in reduced accuracy for either the real or the Galactic scoring model. This is because many of the SHERLOCK features used to create those flags are already given to the model (such as the separation in arcseconds), and the main feature that is omitted (whether the cross-matched source is a galaxy) is one that causes confusion and leads to an overabundance of Galactic sources being flagged as SN. These features therefore either represent a source of confusion or a duplication of information; as a result the (H)GBDT algorithm learns to ignore them.

We note that a low score on the permutation importance graphs shown in Figure 8 can also result when a feature is only available in a small portion of our samples, such as the redshift. It is obvious from an astrophysical standpoint that redshift is a useful quantity, but since it is only available for

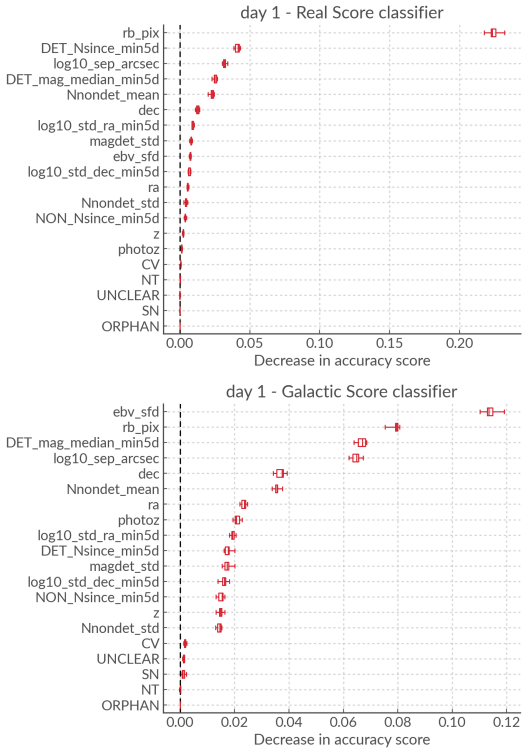

Figure 8. Permutation importance of the day 1 model features.

roughly 13% of our alerts, a metric like the permutation importance does not necessarily reflect how informative that feature is for those samples.

That is why before pruning features we must verify that omitting them does not affect our science metric, in this case the R@K. In Figure 7 we plot both the R@K curves for the models trained on all features (black) and those trained omitting the SN, UNCLEAR, ORPHAN, and NT categories (red). As we can see the curves are virtually identical, confirming that these features are not informative and can be removed from the training process.

3.3.3. *Day* N *Models*

For the models that update the scores when new lightcurve information is recorded (whether detections or nondetections), we need a data set with samples of varying lightcurve completeness. To create this sample we take all the alerts in our training and validation sets for the day 1 models, and we consider their lightcurve between day 2 and day 15. For each ATLAS visit in that time range, we calculate the new lightcurve features described in Section 3.2.3 (also see Table 1). This means that a single object will be represented several times in this data set, increasing the number and distribution of alert types in our training and validation sets, as we can see in Table 3. The class balance is a little skewed toward the "PM"-labeled data but overall it is sufficiently balanced that we do not consider resampling.

In the left panel of Figure 9 we show the score space for all the predictions on our validation set. The $p_{\rm gal}$ scorer performs better in the day N models than in the day 1 models with steeper inclines for the "good" and "Galactic" distributions. Although the "garbage" and "PM" distributions look like they stretch across most of the 2D space, we can see when looking

Table 3
Day N Model Training and Validation Set Label Distribution

Label	Training	Val.
Good [$p_{\rm real} = 1$; $p_{\rm gal} = 0$]	18,798	3331
Galactic [$p_{\rm real} = 1$; $p_{\rm gal} = 1$]	17,695	2972
PM [$p_{\rm real} = 0$; $p_{\rm gal} = 1$]	15,655	5537
Garbage [$p_{\rm real} = {\rm NaN}$; $p_{\rm gal} = 0$]	14,812	31,393
Auto-garbage [$p_{\rm real} = {\rm NaN}$; $p_{\rm gal} = 0$]	11,711	26,262
TOTAL	78,671	69,495

at the marginalized $p_{\rm real}$ distribution that they are very similar to those in the day 1 score space plot in Figure 7, but due to the larger number of samples even with medium transparency the scatter plots are inevitably crowded. We chose to plot individual points rather than, say, a kernel density approximation because these would hide the outliers in the "good" and "Galactic" classes that are found in unexpected areas of the plot and are informative during development.

In the right panel of Figure 9 we can see the R@K curve for the day N models. The R@K plot here is not as directly interpretable as that of the day 1 models, because all samples are ordered together and not separated by visits or day N.[13] Separating by visits would not be a useful comparison as the eyeball list on any given day is mostly composed of new targets. Therefore, sorting a handful of objects that are all on their third visit does not tell us how these rankings would perform in a complete eyeball list. A more direct evaluation of how our models and our eyeballing policies perform together can be found in Section 4.1. The main takeaway from the R@K plot for the day N models is the comparison to the day 1 models: If they did not perform better on the whole than the day 1 models, it would mean that the features we have extracted on the new lightcurve data are uninformative. On the contrary we see that the day N models substantially outperform the day 1 models in placing "good" objects at the top of the list, with a 95% (99%) recall achieved in the top 5% (20%) of the list. Their AuRaK is 0.966.

This indicates that the new lightcurve features are helping with the classification and that having a separate type of model that updates the scores after an object has already entered the eyeball list is useful.

We can also calculate the permutation importance of each feature (also using 10 repeats)—see Figure 10—for which we include the SHERLOCK flags to check that, for the day 1 models, they do not provide new or useful information compared to the other features. Based on Figure 10 we prune SN, UNCLEAR, ORPHAN, and NT, as well as `DET_N_today` and `NON_N_today`. The pruned models perform equally well (see right panel of Figure 9) and these features are removed from the final models. The low importance of `DET_N_today` and `NON_N_today` can be once again interpreted as the result of features duplicating information. `DET_N_today` and `NON_N_today` are the number of detections and nondetections (respectively) seen on a particular visit, but since we already calculate the total number of detections and nondetections seen in total since −5 days (see Table 1 for the full list of features and their descriptions), this information is already included in the total count. Although `DET_mag_median` and `NON_mag_median`—the median

[13] Since our cadence is not 24 hr visit N and day N are not the same value.

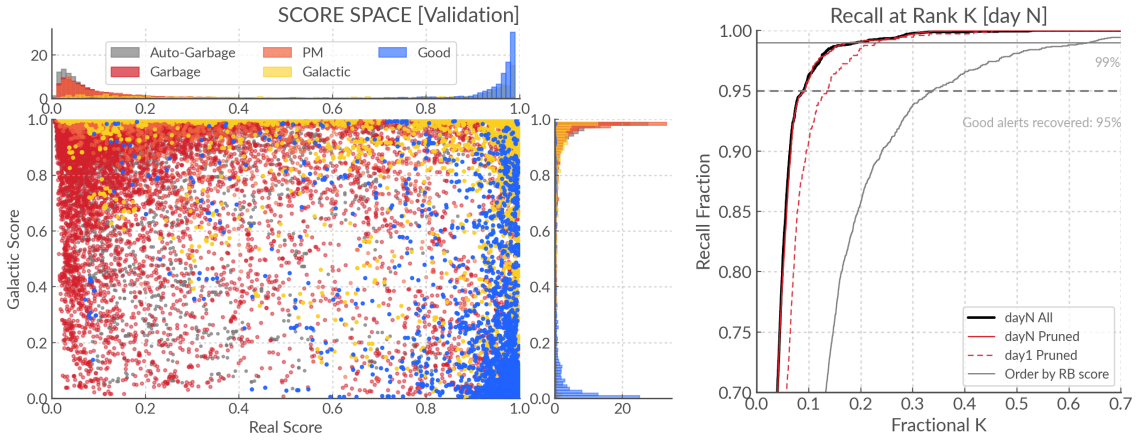

Figure 9. Score space and R@K of our validation data set for the day *N* models, color-coded according to their human-given labels. The score space is that obtained with the models trained excluding the pruned features. The R@K plot shows both the R@K curves obtained for the day *N* models (solid lines) trained on all the features (black line) and those trained excluding the pruned features (red line). We also show the R@K curves obtained with our day 1 model (dashed line). For comparison with our benchmark in Figure 3 we show the R@K curve obtained when ordering by the RB scores calculated by the CNN.

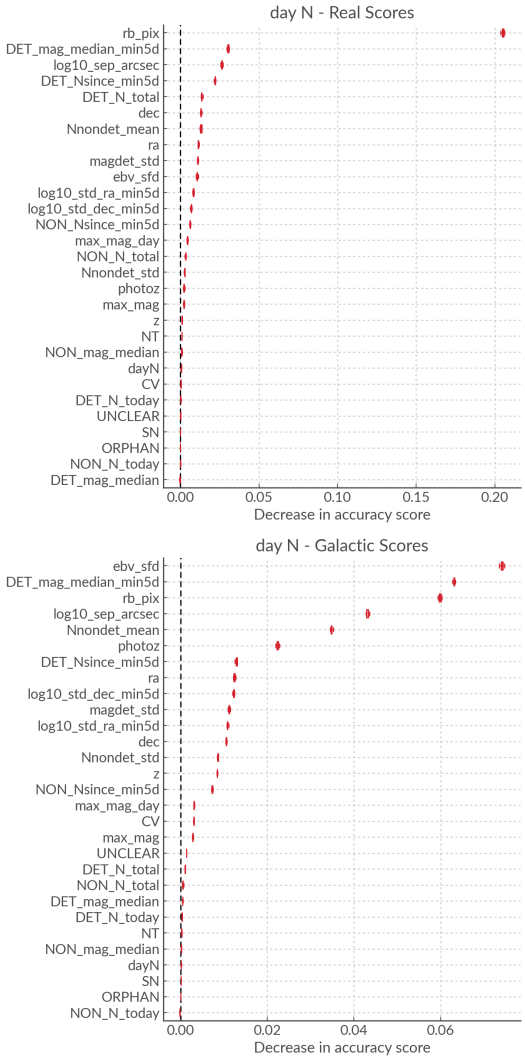

Figure 10. Permutation importance for the day *N* models.

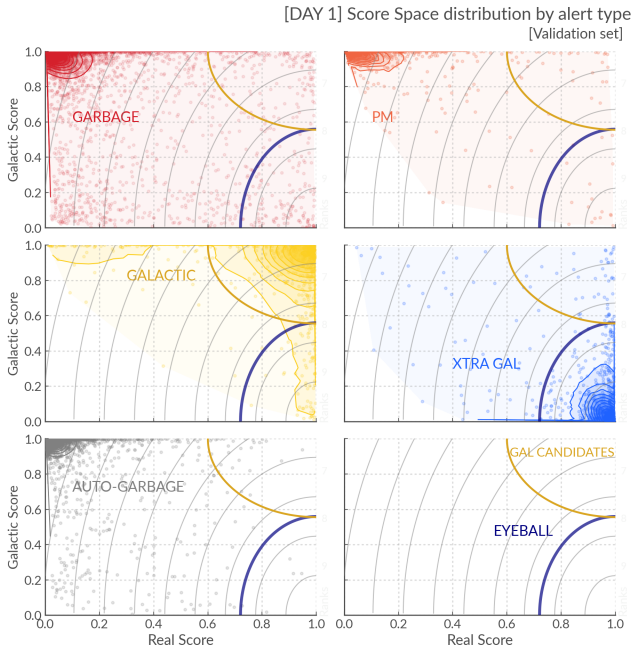

Figure 11. Visualization of the extragalactic eyeballing policies (blue line) and Galactic candidate policies (yellow line) compared to the distribution of our alert samples. All VRA ranks from 1 to 9 are also shown (gray solid lines).

magnitude of the detections and nondetections on the day—are lower in permutation importance, we find that removing them leads to more change in the R@K curve and a longer tail in the "good" object $p_{\rm gal}$ distribution so we do not prune these features.

For HGBDT models, pruning is not strictly necessary (performance will not decrease if uninformative features are included), but we choose to remove the most uninformative features to limit how much data processing has to be done in production when running the VRA to limit code bloat.

4. VRA Performance

As described in Sections 2.4 and 2.5, we have policies to rank the most promising extragalactic candidates and move the most likely Galactic transients into a separate eyeball list and auto-garbaging policies to clear the bulk of the bogus detections. We can visualize these policies against our model predictions in Figure 11. The ranking strategy has already been

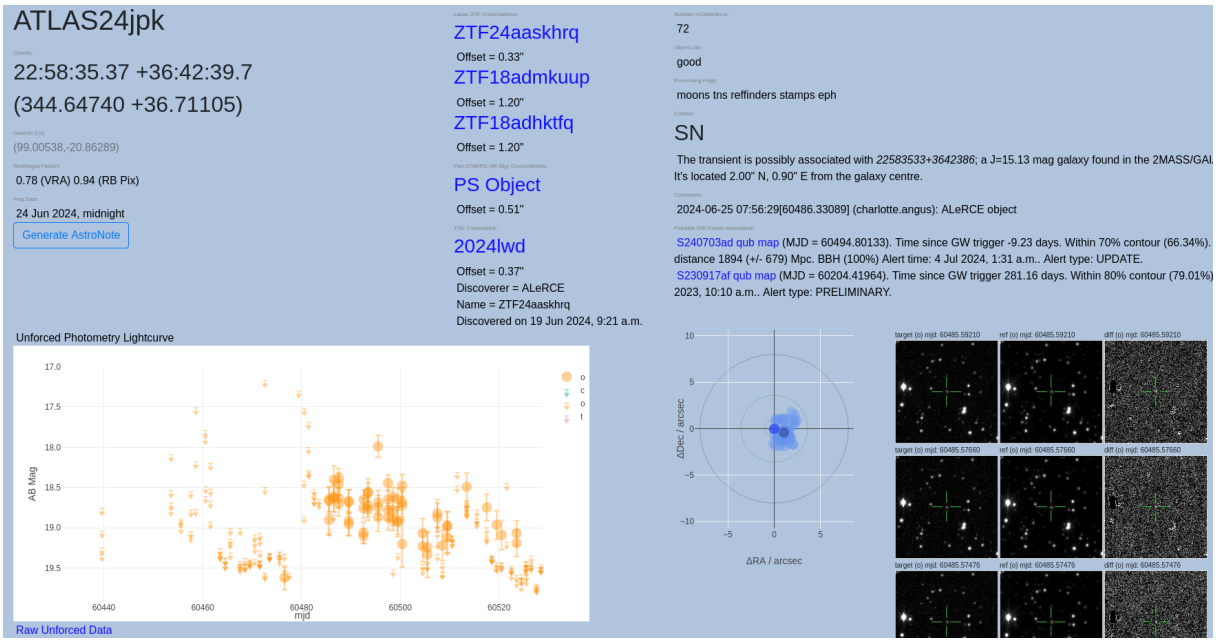

Figure 12. Transient web server page for AT 2024lwd, which is mistakenly labeled as garbage by the VRA day 1 models during policy evaluation.

evaluated using the R@K curve and AuRaK, but the threshold below which human scanners are not asked to eyeball has not been considered. In this section we evaluate how the models and policies interact with each other before implementing a model in production.

4.1. Policy Evaluation

4.1.1. Day 1

First we apply the rankings and policies described in Section 2 on our day 1 validation data set. We find that 90% of the "good" objects are eyeballed on day 1, and only three objects (0.32%) are auto-garbaged. The first is AT 2024ugz, which is very faint and has a lower RB score (0.87). The second is AT 2024aayb and it received a very low RB score (0.28). The final object to be auto-garbaged on day 1 is AT 2024lwd, with $p_{\rm real} = 0.082$ and $p_{\rm gal} = 0.9902$ (for a $\rm VRA_{score} = 0.67$); it is predicted by our classifiers to be a bogus alert correlated with the Galaxy. When inspecting the web server (see Figure 12) it is not immediately obvious why this alert would receive these predictions: the lightcurve is clear, the RB score is quite high (although not 1.0), and the on-sky location is not particularly near the Galactic plane ($b = -20°$).

Because we are using feature-based algorithms, our first port-of-call to understand this misclassification is to investigate the features. We find that five of the features calculated for AT 2024lwd have values that are much more consistent with the Galactic and/or garbage population (`Nnondet_std`, `Nnondet_mean`, `log10_std_ra_min5d`, `log10_std_dec_min5d`, and `ebv_sfd`; see Appendix A, Figure 25).

The lightcurve history features `Nnondet_std` and `Nnondet_mean` are high because the rise of the transient happened to coincide with a waxing Moon, as we can see in Figures 12 and 13. A single detection in the days before the alert was followed by a streak of nondetections as the sky grew brighter faster than the transient. Transients rising with the Moon are not uncommon but a single detection (out of four

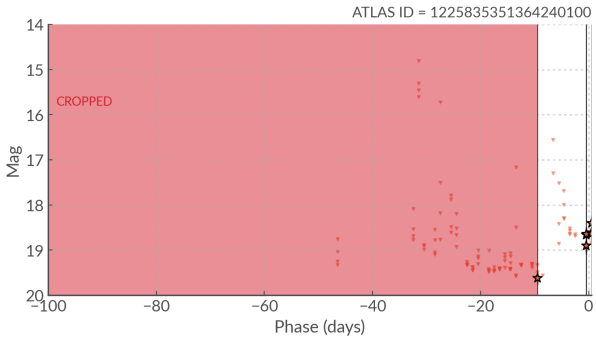

Figure 13. Long-term history of AT 2024lwd (ATLAS_ID = 1225835351364240100) as recorded by the VRA feature calculator.

frames) followed by 10 days of nondetections is a very specific failure mode (at least in our experience—since deploying the VRA this case has not arisen). In combination with an unusually large scatter in the R.A. and decl. measurements, and a sky location that coincides with a slightly elevated extinction, the misclassification as a Galactic bogus alert can be understood as a combination of unlikely events. No action is taken at this stage, but should similar cases be uncovered by cross-matching between the garbage and TNS, we would review which features are most decisive in silencing these alerts and which additional training may need to be performed.

Overall on day 1, 69% of the eyeball list is auto-garbaged, 5% is sent to the Galactic candidate eyeball list, 9% remains in the extragalactic candidate eyeball list, and 18% has been neither eyeballed nor auto-garbaged, awaiting further data. We say that the latter is in purgatory.

For the day 1 models we take an extra evaluation step that consists in looking at how some rare types of extragalactic transients that are particularly relevant to our science team are scored and ranked. These are also an interesting test set since they span several years of ATLAS data going back to 2018. Should data drift be a major issue we should see older objects

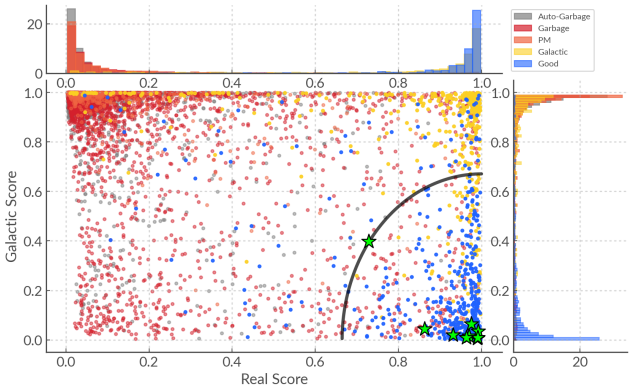

Figure 14. Position of the key transients (lime) listed in Table 4 on the day 1 score space. The extragalactic candidate eyeballing threshold is shown as a black line.

Table 4
Key Transients and Their Day 1 Model Scores and Ranks

Transient	$p_{\rm real}$	$p_{\rm gal}$	$\rm VRA_{score}$	VRA Class
AT 2018cow	0.990	0.006	9.904	Extragal.
AT 2018kzr	0.931	0.019	9.380	Extragal.
SN 2023zaw	0.975	0.064	9.637	Extragal.
SN 2023ufx	0.992	0.035	9.827	Extragal.
AT 2024eju	0.728	0.400	6.987	Purgatory
SN 2024atk	0.991	0.008	9.910	Extragal.
SN 2020kyg	0.970	0.015	9.723	Extragal.
SN 2020aedm	0.963	0.008	9.670	Extragal.
SN 2022ilv	0.863	0.043	8.762	Extragal.

perform significantly worse. The chosen objects are listed in Table 4 and their predictions are shown in score space alongside the validation set in Figure 14. As we can see, all fall well within our eyeball threshold of 7 (see Section 2) except for AT 2024eju, which has a rank 0.03 lower (rank 6.987). AT 2024eju is qualitatively different to the rest of the transients in Table 4: it had a high RB score (0.99), but was only detected on one night and was initially thought to be a Galactic source by a human scanner. However it was the optical counterpart to the fast X-ray transient EP 20240315a as described in J. H. Gillanders et al. (2024). This information from an external source (a 3′ localization radius) allowed us to recognize it as an extragalactic alert. For such a borderline alert to have a $\rm VRA_{score}$ within 0.03 of our eyeballing threshold is a good indication that our policies are robust despite removing nearly 70% of the stream on day 1.

Finally we do not find that older objects perform any worse than more recent ones. Although this is not a sufficient test to claim that data drift will be a negligible issue, it is encouraging.

4.1.2. After Four Visits

To evaluate how the day N models and later auto-garbaging policies perform, we take the samples that are left in purgatory and apply the corresponding policies before once again separating the alerts into extragalactic eyeballing, Galactic eyeballing, and purgatory. We repeat these successive steps for a total of four visits, which given the ATLAS cadence corresponds to between 4 and 15 days (our cutoff) after the initial alert on average.

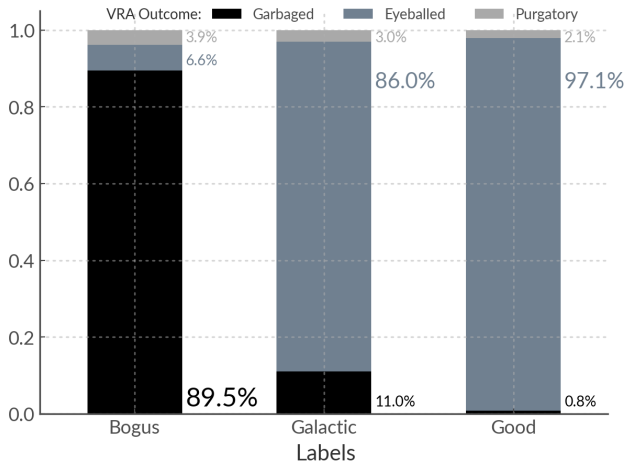

Figure 15. VRA outcome for the "good," "Galactic," and "bogus" alerts (which include "garbage," "PM," and "auto-garbage") after four visits. The eyeballed outcome encompasses both alerts that are eyeballed as extragalactic and Galactic candidates.

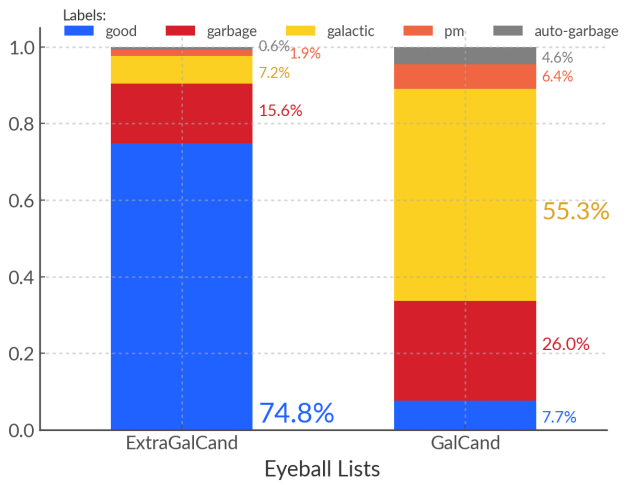

Figure 16. Mixture of alert types in the extragalactic candidate and Galactic candidate eyeball lists (total after four visits).

We can see in Figure 15 the fate of our data split by type (combining the "auto-garbage," "garbage," and "PM" categories into one bin). We can see that 97% of the good objects are eyeballed (either through the extragalactic or the Galactic eyeball list). The 2% that remain in purgatory are checked by eye. Many are rather faint sources that were found by other surveys first and would have been eyeballed in production because their $\rm VRA_{score}$ would have been automatically raised to 10. All are alerts that we are happy can be eyeballed with a delay of 15 days to add to the good list for completeness: none would have been high-priority follow-up targets where a 15 day latency would mean a missed opportunity. More details can be found in the code release, particularly in the "Policy_-evaluation" Jupyter notebook. Additionally we have a good recovery of Galactic transient events, with only 11% being discarded.

Another important consideration is the composition of the extragalactic and Galactic eyeball lists. As we can see in Figure 16, the cumulative extragalactic eyeball list (after four visits) is composed of >80% real transients, 75% of those

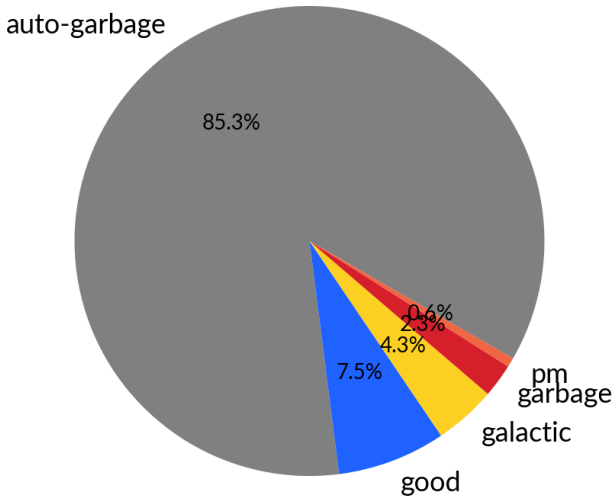

Figure 17. Alert type distribution over the period 2025 April 4 to 2025 June 10. See the label description in Section 3.1.

being extragalactic objects we are targeting. The Galactic candidate list contains a higher fraction of bogus alerts but still <40%. As for the "good" objects in the Galactic eyeball list they were visually inspected and many were mislabeled CVs or interesting fast extragalactic transients that look similar.

Overall we consider our policies and how they interact with the models to be satisfactory. Based on our validation set, we can expect that over the course of a week 80.2% of the incoming alerts will be auto-garbaged, 3.7% will be left in purgatory, and 16% will be sent to human scanners for eyeballing. Of the 3.7% left in purgatory during our tests 16 were labeled as "Galactic" and 14 as "good" objects. As mentioned above we visually checked those 14 objects and found no concerning outliers.

4.2. In-production Performance

We now present the performance of the VRA between 2025 April 4 and 2025 June 10, during which 16,938 alerts entered the eyeball list. As we can see in Figure 17 the VRA auto-garbaged 85% of the alerts over that period. Our policy evaluation estimated that 80.2% of objects would be handled by the VRA; the slightly better in-production performance is the result of acute hardware or weather events that were not taken into account in our tests. In this case some ATLAS units were subject to significant trailing in the images on the week starting 2025 May 23 leading to 7516 alerts entering the eyeball list on that week alone. During such an event the VRA auto-garbages a higher fraction of alerts (91.5% on that particular week), raising the average for the month.

The TNS cross-match to the garbage over this period found transient events mislabeled as garbage but these were the result of human error. There were 164 "potential misses"—alerts that did not meet the VRA score threshold initially but whose VRA score was raised to 10 when a cross-match to TNS was detected. Of these, 34 would not have risen above our threshold within the 15 day period covered by the models. In eyeballing these 34 events, we found that 11 were duplicates, two were CVs, and one was suspected to be bogus. Of the 20 real events left, only one may have been the object of follow-up—SN 2025hkm. The VRA scores for that event rose and stabilized to about 6.6 for over a week, still below our threshold likely due to a slightly low RB score (0.67).

This indicates that future versions of the VRA could benefit from an additional feature that specifically counts the number of successive detections, or potentially for objects with VRA scores above 6 for a few days in a row we could trigger the CNN again to update the RB score as the object is now brighter.

We leave this for future enhancements, as the current potential loss rate is only 0.006% of all the alerts entering the eyeball list, and 0.079% of all the extragalactic events.

5. Discussion

5.1. Raw versus Forced Photometry

All of our lightcurve features are based on the raw (difference) photometry measured on the night. As a result we have to handle a mixture of detections and nondetections, and we are vulnerable to the effects of the waning and waxing Moon (which affects our detection limits and can turn a detection into a nondetection). An obvious solution to this problem would be to use the forced photometry lightcurves instead. There is however a technical limitation and computational cost in production, which makes this unfeasible at this stage. The forced photometry is not calculated for ATLAS alerts entering the eyeball list unless they meet specific quality criteria. This is to ensure that the load on the computer servers results in the highest-priority transients being processed fast and the forced photometry speed for those is not compromised. Consequently, at the point at which the VRA runs in the stream, forced photometry is not available in the majority of cases, and our algorithms are trained on the data readily available in production.

We use this opportunity to highlight that this is an example of "data first" design in applied ML. The growing literature on ML in astronomy often suffers from a "model first" approach, where models are applied to a problem with no clear benchmark of success and the data provided to the algorithms is optimistic (if not at times unrealistic) compared to what we would expect in real-life settings. This is where proofs of concept can fail to lead to practical science solutions, because the limitations of the data ignored at the design stage are not easily overcome.

Focusing back on the case of the VRA, its primary job is to reduce the false-positive rate in the data flow without causing time delays or loss of opportunity. A next iteration would be to explore additional features that harness the forced photometry in the day N models since 70% of the stream is cleaned on day 1, lowering the cost to compute the forced photometry. This would allow analytic fits, extracting gradients, and model comparison.

5.2. The Fast Axis

Related to the discussion of using unforced photometry of detections or forced photometry at a known position is that of scoring alerts on the fast axis imagined in Figure 4. One way to create transient-agnostic scoring for this axis would be to evaluate the gradient of the lightcurve and apply a normalization factor. Without the forced photometry, the calculation

of this gradient is compromised by the uncertainty introduced by nondetections.

There are other simpler alternatives such as using the max_mag feature or creating a Boolean flag where all alerts with day 1 magnitude below, say, $m = 16$ are considered fast. These methods are strongly biased toward the CV population but could highlight nearby FBOTs such as AT 2018cow, or shock breakout events. AT 2018cow was discovered and recognized due to its very rapid rise to bright absolute magnitude (S. J. Prentice et al. 2018; S. J. Smartt et al. 2018).

On the whole since what is considered "fast" is a less objective classification than real or Galactic, and since creating a useful transient-agnostic scoring method would not be feasible with the data available at the time of scoring, we omitted the fast axis from our score space for this specific use case. In larger streams (e.g., LSST or future surveys) having an additional dimension to rank and prioritize alerts could provide sufficient benefits to justify adding complexity to the ranking method.

5.3. Feature Importance

In Figures 8 and 10 we showed the permutation importance of our features for the day 1 and day N models. In this section we highlight a few takeaways from this analysis.[14]

The first notable feature is rb_pix (RB score). As expected it has the largest influence on the $p_{\rm real}$ score and, perhaps surprisingly, the second (third) largest on the $p_{\rm gal}$ scores of the day 1 (day N) models. The relation between rb_pix and alert type is clear when we look at the rb_pix distribution by type (see Figure 18 in Appendix A).

Another key feature is the extinction (ebv_sfd), which is the most predictive feature for the $p_{\rm gal}$ scores. This is not surprising since high extinction is correlated with the plane of the Galaxy, where more Galactic events may occur, and makes extragalactic transients fainter and less likely to be observed.

Since a correlation with the Galactic plane is important in determining whether an alert is Galactic or extragalactic, it may seem surprising that we do not use Galactic coordinates for our scoring algorithm. In fact in earlier prototypes of the VRA we did test using the Galactic latitude as a feature instead of R.A. and decl. but found that our $p_{\rm gal}$ classifier performed much worse (see H. Stevance 2025a). There are two reasons for this: as we can see in Figure 19 in Appendix A there is a strong correlation between the Galactic plane and the bogus alerts; then there is the effect of the Galactic center, which can only be accounted for with 2D coordinates (R.A. and decl. or Galactic latitude and longitude). At the time of those tests the conclusion was to keep R.A. and decl. coordinates as they were already in the stream, and try including $E(B - V)$ as a feature that had not yet been tested, which was then found to be very informative.

The final features we will discuss are those related to redshift, z and photoz, as we find the results from the permutation importance surprising. We expected z to be consistently more important than photoz—since spectroscopic measurements of redshift are more reliable than photometric measurements—but found the converse. It could have been due to the larger availability of photometric redshift measurements; however we find in our data that 2905 samples have spectroscopic measurements, and 2748 have a photo-z, so greater availability of one measurement over the other is ruled out as a cause. We do not have a firm explanation for this discrepancy but we put forward a few hypotheses that could be tested in a future version. A first possibility is that the value of redshift is less important than the fact that there is a redshift at all—we could test this by turning the photoz and z features into Boolean flags (individually and then together). A second possibility is that the relevance of the redshift split points in the decision trees is minimally affected by the errors on the redshift. Testing this is more difficult; we could try artificially adding noise to the photoz and z features and see how this affects their position in the permutation analysis.

5.4. Choosing the Policies

Calculating the VRA score and Galactic flag and applying our eyeballing and garbaging policies requires seven values to be set: two scalar values f for Equation (2) (one used when calculating the VRA score ($f = 0.5$), and one used when calculating the Galactic flag ($f = 0.9$)); an extragalactic candidate eyeballing threshold (>7); a distance to the (1, 1) coordinates in score space to set the Galactic flag to "true" (<0.4); and currently three auto-garbaging policies for the first, second, and third visits.

For our calculation of the $\rm VRA_{score}$ in this version of the VRA we tested f values ranging from 0.4 to 1.[15] We found that values of 0.4, 0.5, and 0.6 give very similar results and all have an AuRaK of 0.951.[16] We chose $f = 0.5$ because it has the convenient interpretation of weighing the "real" axis twice as much as the "Galactic" axis.

To calculate the Galactic flag, the choice of $f = 0.9$ and distance <0.4 did not go through a systematic search. Instead, the values were chosen to conservatively cover the Galactic distribution without encroaching too much on the bogus distributions based on the visualizations in Figure 11. One could create a larger grid search for the extragalactic and Galactic policy values and rerun all our policy diagnostics to optimize these values. This was not done because from the earlier grid searches we noted that optimized parameters are usually very close to ones chosen by visual inspection of the plots, and given the current performance such a systematic search of parameters was not considered necessary. In the future this may take place as we review the live performance of the current VRA over the next few months.

Finally the auto-garbaging policies were created during the first live implementation of the VRA in 2024 August. The in-production VRA scores were recorded for all alerts in the eyeball list and the objects were eyeballed as usual. This provided us with a first real test set, and the distribution of the VRA scores in this set was used to establish conservative eyeballing policies one after another. In future iterations of the VRA the garbaging logic remained the same but the values changed slightly based on the policy evaluation presented in Section 4.1. For a history of the garbaging policies see the technical manual (H. Stevance 2025a).

[14] However, note that looking at individual features is an incomplete form of interpretation because the models use combinations of features to make decisions.

[15] Tests on earlier VRA prototypes with $f = 0.1$, 0.25, 0.5, 0.75, 1.0 can be found in H. Stevance (2025a).

[16] See code release (H. Stevance 2025b).

6. Summary and Conclusions

The ATLAS VRA is a bot that performs preliminary eyeballing to rank and prioritize alerts for human eyeballers. It has reduced eyeballing workload by 85% with no loss of follow-up opportunity.

It uses histogram-based gradient-boosted classifiers to predict a "real" ($p_{\rm real}$) and a "Galactic" score ($p_{\rm gal}$) for each alert and the scores are updated after each new visit by the ATLAS telescopes, up to 15 days after first entering the eyeball list. The $p_{\rm real}$ and $p_{\rm gal}$ values are then used to calculate the VRA score (see Equation (2)), which ranges from 0 (bogus) to 10 (real and extragalactic). We also calculate a "Galactic" flag based on the distance to the $p_{\rm real} = 1$, $p_{\rm gal} = 1$ coordinates. Auto-garbaging policies are applied to remove the alerts most likely to be bogus from the eyeball list, and a VRA score threshold is used to select the alerts to be visually inspected by our team.

VRAs with similar strategies could be very useful to other sky surveys such as GOTO or BlackGem (P. J. Groot et al. 2024) to limit the reliance on citizen scientists or offer volunteers classification tasks that are more rewarding and engaging. Although the ATLAS VRA should not be run as is on data from another survey, transfer learning techniques could be worth exploring (H. Domínguez Sánchez et al. 2019; R. Gupta et al. 2025) so that the VRAs of other teams can be trained using smaller training sets. There may be issues with the differences in survey cadence and magnitude limits affecting the lightcurve features, and the strong dependence on the `rb_pix` value that is specific to our RB classifier. If these effects render transfer learning impossible, the advantage of the VRA design is that a relatively small sample (a few thousand) can provide good results in production.

The success of the ATLAS VRA demonstrates that our field has not fully leveraged the potential of feature-based ML methods, and we encourage our colleagues to not dismiss these without experimentation as they provide several advantages. First, they can be trained with only a few thousand (sometimes a few hundred) samples, which means that we did not have to rely on synthetic data. Additionally feature-based methods are easier to interpret and provide us with a direct way to inject our expertise into the models (for a discussion see Appendix B). (H)GBDT in particular has native support for categorical features and null values (without imputing).

Finally, the performance of the VRA has allowed us to introduce (since 2024 December) an automated trigger mechanism for the 1 m Lesedi telescope and the Mookodi instrument (H. L. Worters et al. 2016; N. Erasmus et al. 2024b), as part of the South African Astronomical Observatory's "Intelligent Observatory" (N. Erasmus et al. 2024a; S. B. Potter et al. 2024). Automated triggers have already resulted in classification (e.g., SN 2025arc[17]), and our criteria are still being refined to increase the number of eligible alerts while minimizing unnecessary trigger. These tests are important precursors to the automated triggering system that needs to be deployed on the LSST stream to shorten follow-up latency on instruments such as SOXS (K. K. Radhakrishnan Santhakumari et al. 2024).

Our next focus will be to adapt the ATLAS VRA to data brokers such as Lasair and Fink (A. Möller et al. 2021; R. D. Williams et al. 2024). We welcome discussions and collaboration from other survey teams should they wish to use our design to curate their data stream.

[17] https://www.wis-tns.org/object/2025arc

Acknowledgments

We are grateful to the referee for very helpful comments and suggestions. H.F.S. is supported by Schmidt Science. K.W.S. and S.J.S. are supported by the Royal Society. This work has made use of data from the Asteroid Terrestrial-impact Last Alert System (ATLAS) project. The Asteroid Terrestrial-impact Last Alert System (ATLAS) project is primarily funded to search for near-Earth asteroids through NASA grants NN12AR55G, 80NSSC18K0284, and 80NSSC18K1575; byproducts of the NEO search include images and catalogs from the survey area. This work was partially funded by Kepler/K2 grant J1944/80NSSC19K0112 and HST GO-15889, and STFC grants ST/T000198/1 and ST/S006109/1. The ATLAS science products have been made possible through the contributions of the University of Hawaii Institute for Astronomy, the Queen's University Belfast, the Space Telescope Science Institute, the South African Astronomical Observatory, and the Millennium Institute of Astrophysics (MAS), Chile. This work made use of observations made at the South African Astronomical Observatory (SAAO), which is supported by the South African National Research Foundation.

Facilities: ATLAS, SAAO:1m.

Software: atlasapiclient (H. F. Stevance et al. 2025), atlasvras (H. Stevance & K. Smith 2025), pandas (W. McKinney 2010; The pandas development team 2020), numpy (C. R. Harris et al. 2020), sklearn (L. Buitinck et al. 2013), matplotlib (J. D. Hunter 2007).

Appendix A
Feature Distributions

In this appendix we show plots of the distributions of all our features split into five types—"good," "Galactic," "PM," "garbage," and "auto-garbage" (as previously defined)—and add supplementary information that is not discussed in the main text.

A.1. Contextual Features

A.1.1. RB Score from the CNN

The RB score from the CNN (`rb_pix` feature) is crucial in predicting the real and Galactic scores (see Section 3.3). In Figure 18 we can see that the lower tail of the distribution is what distinguishes the "good" (and "Galactic") alerts from the bogus alerts ("garbage" and "PM"). Another notable characteristic highlighted in the figure is that the "auto-garbage" alerts do not have a pronounced spike in `rb_pix` at and around 1. This is because alerts with a very high `rb_pix` value are unlikely to meet the auto-garbaging policies as their overall VRA score will not be sufficiently low.

A.1.2. R.A. and Decl.

In Figure 19 we show the R.A. and decl. distribution of our labeled data. We can see that the density of PM and garbage objects is affected by the start of VRA operations in 2024 August, after which fewer objects were labeled in these categories. The alerts that would have been labeled as such by human scanners were for the most part auto-garbaged, which is noticeable in the auto-garbage distribution, which is visibly denser for R.A. observed after 2024 August.

We can also see in the "garbage" and "auto-garbage" maps the effects of the ATLAS tiling pattern. This is a consequence

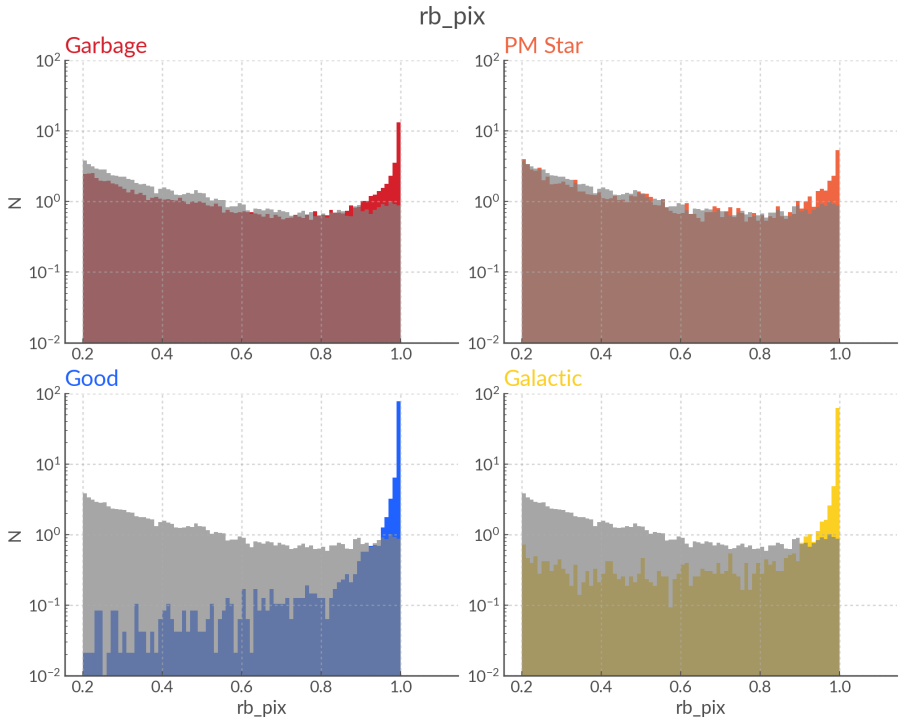

Figure 18. RB score distribution split by alert type. The auto-garbage alerts are plotted in gray over each plot. Note that we logged the *x*-axis for better visualization. The features given to the scoring algorithms are not logged. Also note that the `rb_pix` feature distribution starts at 0.2 because we only use data that passed all previous upstream cuts.

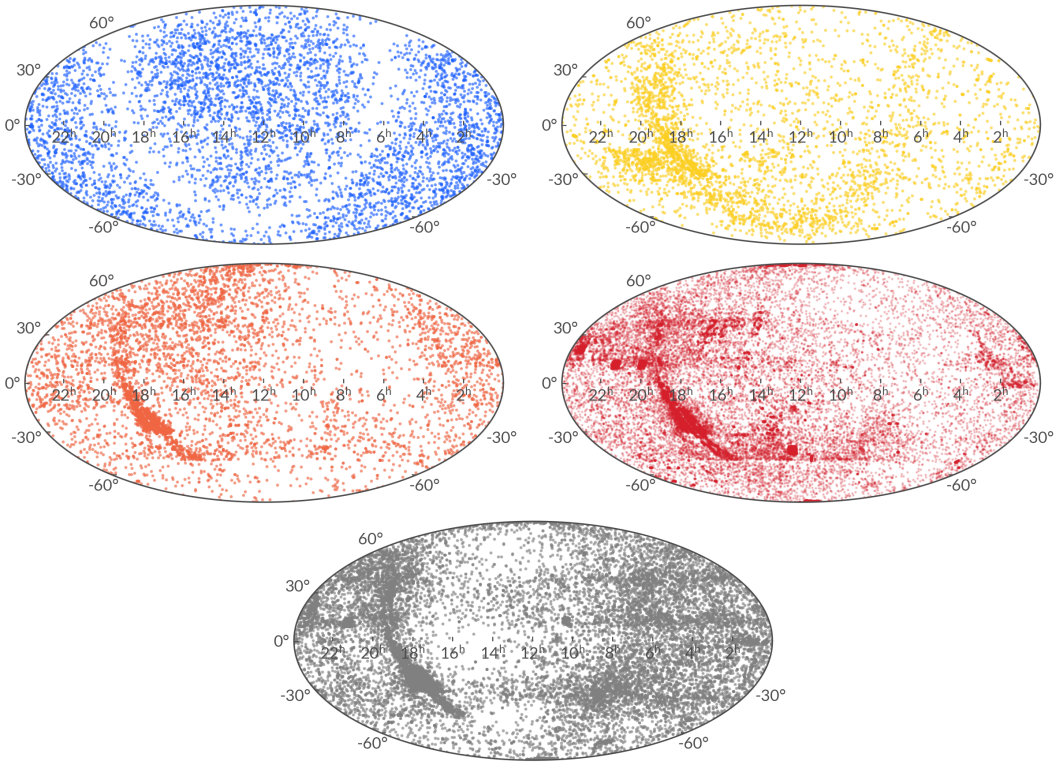

Figure 19. R.A. and decl. distribution of our data split by type: "good" (blue), "Galactic" (yellow), "PM" (orange), "garbage" (red), and "auto-garbage" (gray).

of the higher incidence of bogus alerts on the edges of the ATLAS field of view. Although we have masks to reject alerts from regions that are known to cause problems, these masks are not perfect. Since the tiling pattern is not visible in the map of the "good" alerts (as notably empty lines or patches of sky) we do not think this is an issue, and since introducing the VRA in 2024 August we have not noticed a pattern of missed objects associated with tiling. Since we will continue to monitor

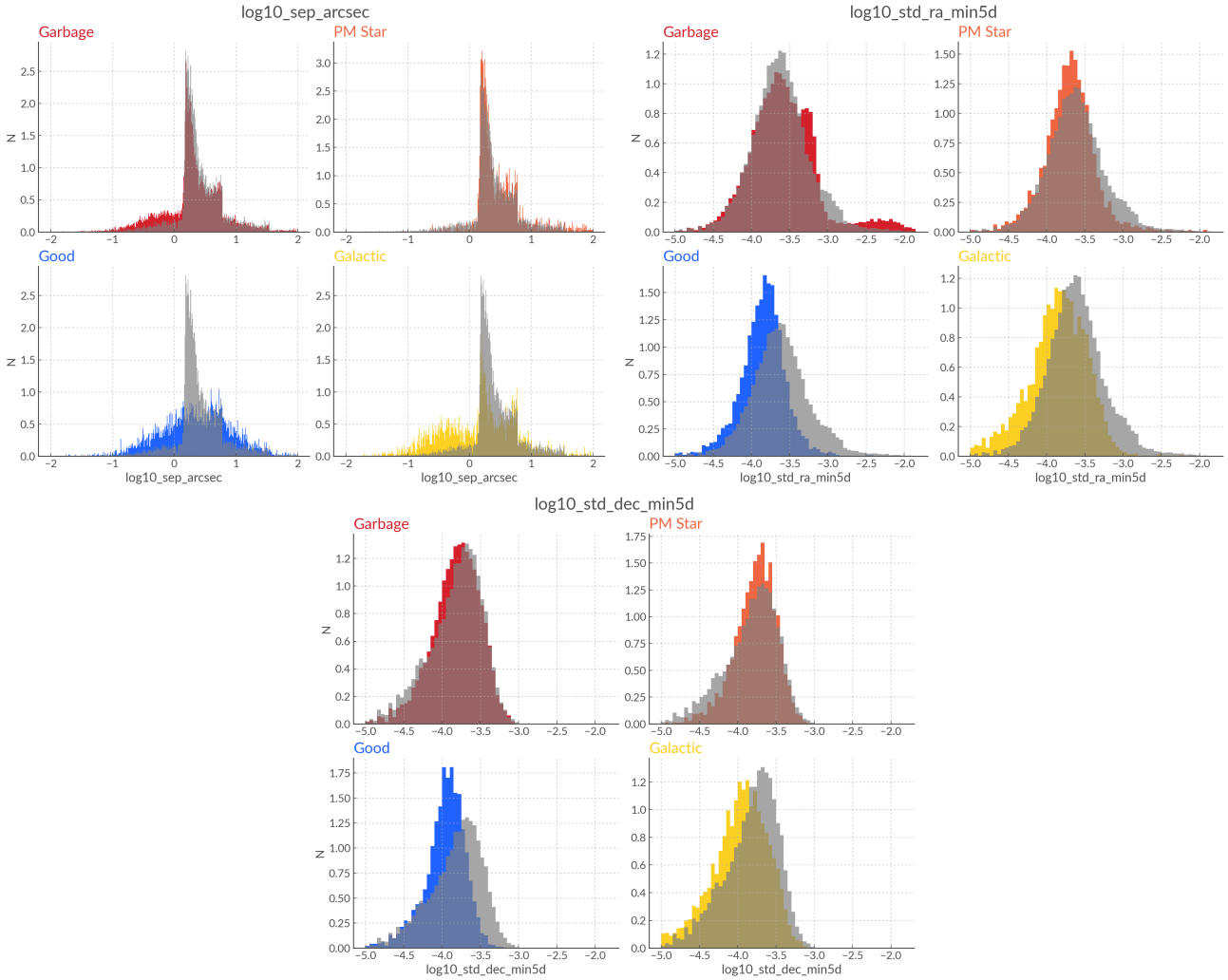

Figure 20. Additional positional feature distributions. (Top left) The separation between the alert and the most nearby cross-matching catalog source; (top right and bottom) the standard deviation of the R.A. and decl. localizations of all detections recorded. All these features were logged (base 10). We plot separately the labels given by human scanners and show the auto-garbage label distribution in gray, overlaid. The auto-garbaged alerts are nearly exclusively garbage and PM stars, which is reflected in how their distributions overlap the other four categories.

potential misses (see Section 2.6) we will assess whether a tiling pattern emerges in our missed transients. If so we can mitigate this by resampling the "garbage" and "auto-garbage" objects across the x, y position on the detectors (balancing the sample across R.A. and decl. would erase the important correlation with the Galactic plane).

In the "garbage" and "auto-garbage" (and PM to a lesser extent) distributions we can also see the decl. limits of the Northern and Southern units and where they overlap. As with the tiling pattern we do not think this is an issue but we will monitor in the long term.

A.1.3. Additional On-sky Localization Features

There are three other features related to the on-sky localization of each alert (see Figure 20): the scatter in R.A. and decl. for detections related to this alert, and the separation of the alert from the catalog source it is associated with (in arcseconds). The latter is provided by SHERLOCK (D. R. Young 2023). All these features were logged (base 10) to increase their dynamic range.

The separation can also be none when there is no viable catalog cross-match. Null values are given as such in our models since the chosen algorithm natively handles null values by reserving one bin of the histogram for them. This is relevant particularly in the case of the separation feature as the null value there represents what SHERLOCK would flag as an orphan detection.

A.1.4. Redshift

The redshift information is known for 4737 alerts or 13.7% of our data set (2905 have a spectroscopic redshift, and 2748 have a photometric redshift). This is provided by SHERLOCK (D. R. Young 2023) if there is an associated catalog source and that source has a known redshift. There are 916 sources for which both the spectroscopic and photometric redshift are known. In Figure 21 we show the feature distribution separated by alert type. We have not taken the extra step of combining spectroscopic and photometric redshift into a single column because we want to keep these two sources of information distinct. The spectroscopic redshift is much more reliable than the photometric redshift and although we cannot easily and

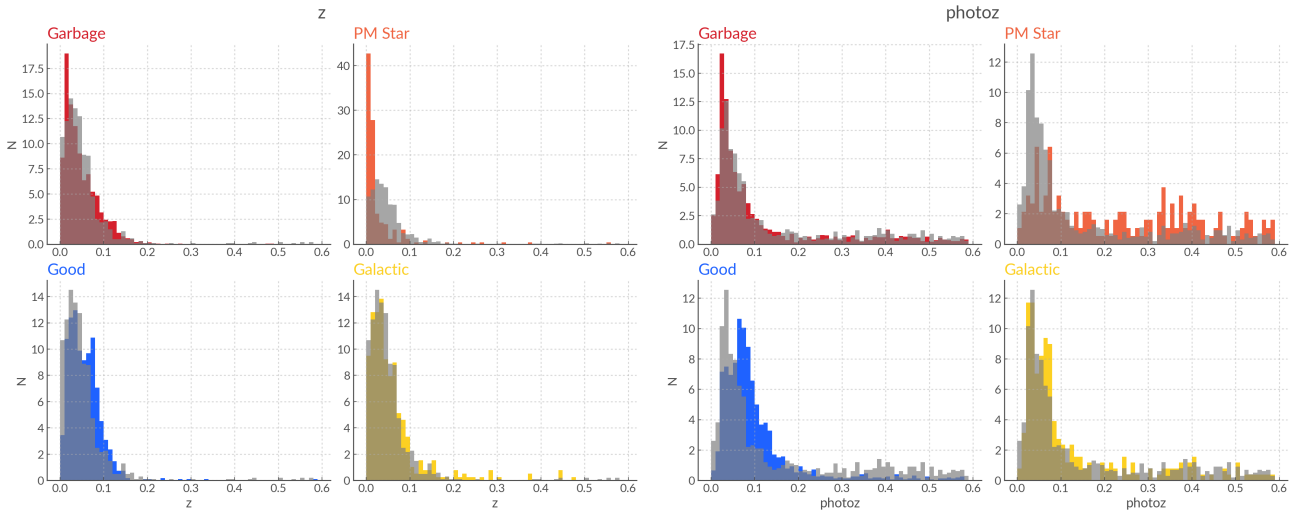

Figure 21. Distribution of the redshift measurements across our alert types. We separate the spectroscopic (z) and photometric redshifts (photo-z). We plot separately the labels given by human scanners and show the auto-garbage label distribution in gray, overlaid.

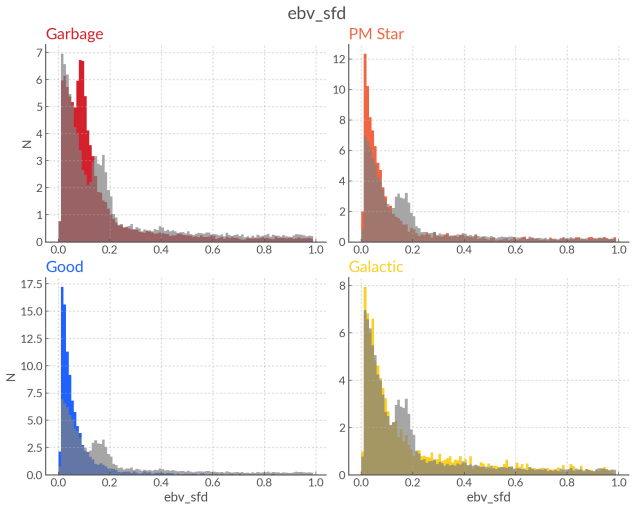

Figure 22. Distribution of the $E(B-V)$ feature for our different alert types. We plot separately the labels given by human scanners and show the auto-garbage label distribution in gray, overlaid.

formally "tell" the models, by keeping these two pieces of information separate the decision trees can learn to use one over the other.

A.1.5. Galactic Extinction: E(B − V)

The Galactic extinction feature is calculated using the `dustmaps` Python package by G. Green (2018) and selecting the D. J. Schlegel et al. (1998) extinction maps. In Figure 22 we show the $E(B-V)$ distribution separated by alert type. Unlike previous plots where the auto-garbage distribution shows a behavior very similar to the garbage and PM labels, in this case there is a secondary peak around $E(B-V)$ values of 0.2 that is not shown in any of the other distributions. We interpret this behavior as follows: There is a secondary peak in the garbage $E(B-V)$ distribution around 0.1; however the fraction of good and Galactic (therefore real) objects with $E(B-V) \approx 0.1$ is still significant. Therefore the auto-garbage distribution does not exactly follow that of the garbage alerts. The secondary peak remains but has a lower amplitude and moves to ≈0.2, where the fraction of good and Galactic alerts is lower.

A.2. Day 1 Lightcurve Features

A.2.1. Long-term History

For a description of how these features are calculated see Section 3.2 and Figure 6. In Figure 23 we show the distributions of the three long-term history features.

A.2.2. Short-term History

See Section 3.2 for the description and motivation of the two short-term lightcurve history features. We show the distribution of these features in Figure 24.

A.3. Day 1 Features of AT 2024lwd

We only show the features that are most out-of-distribution for a "good" alert.

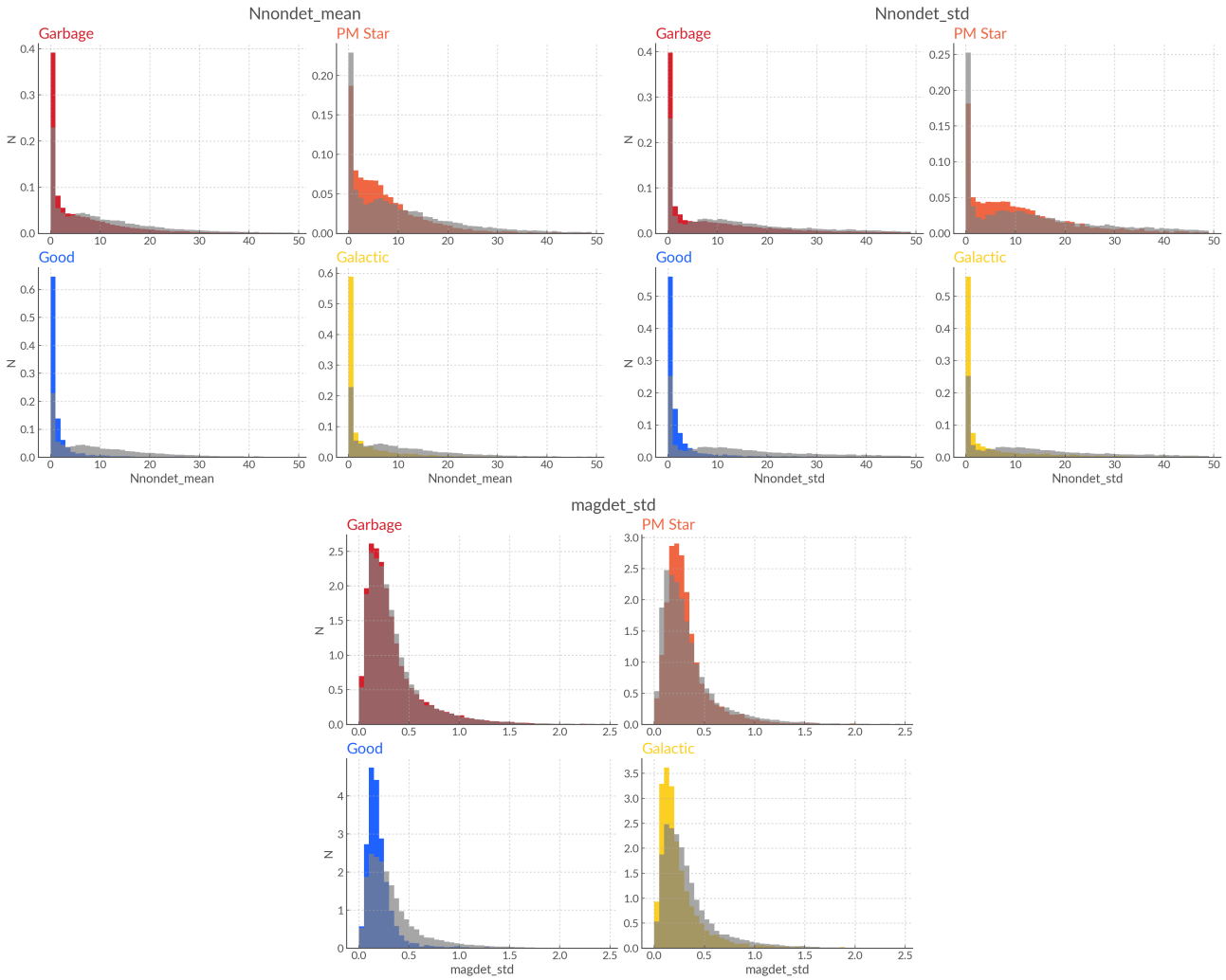

Figure 23. Long-term history (from −100 days with respect to first alert) features, from top to bottom: the mean and standard deviation of the number of nondetections between each detection, and the standard deviation of the magnitude values of these detections. We plot separately the labels given by human scanners and show the auto-garbage label distribution in gray, overlaid.

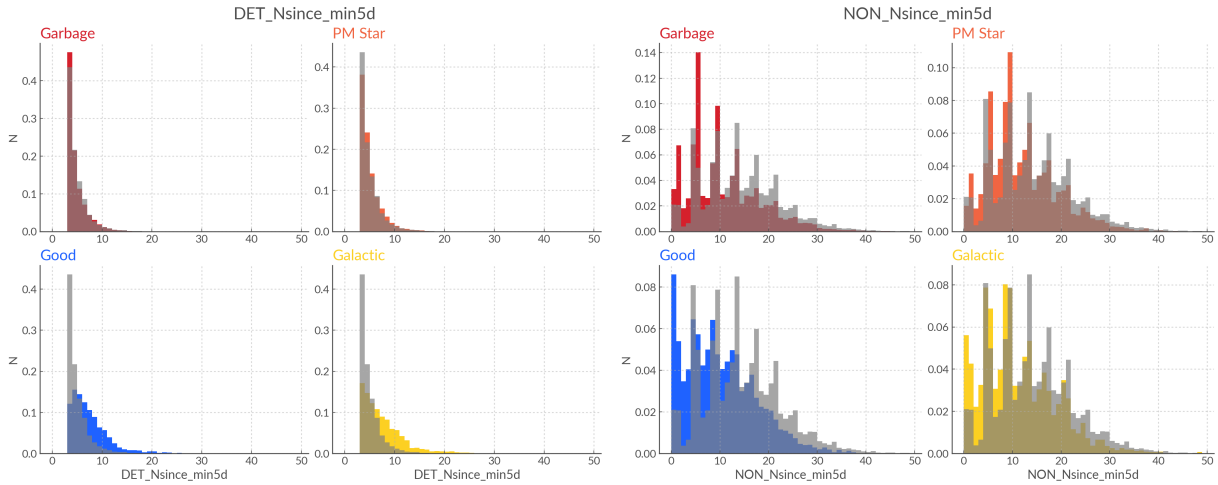

Figure 24. Short-term history features. We plot separately the labels given by human scanners and show the auto-garbage label distribution in gray, overlaid.

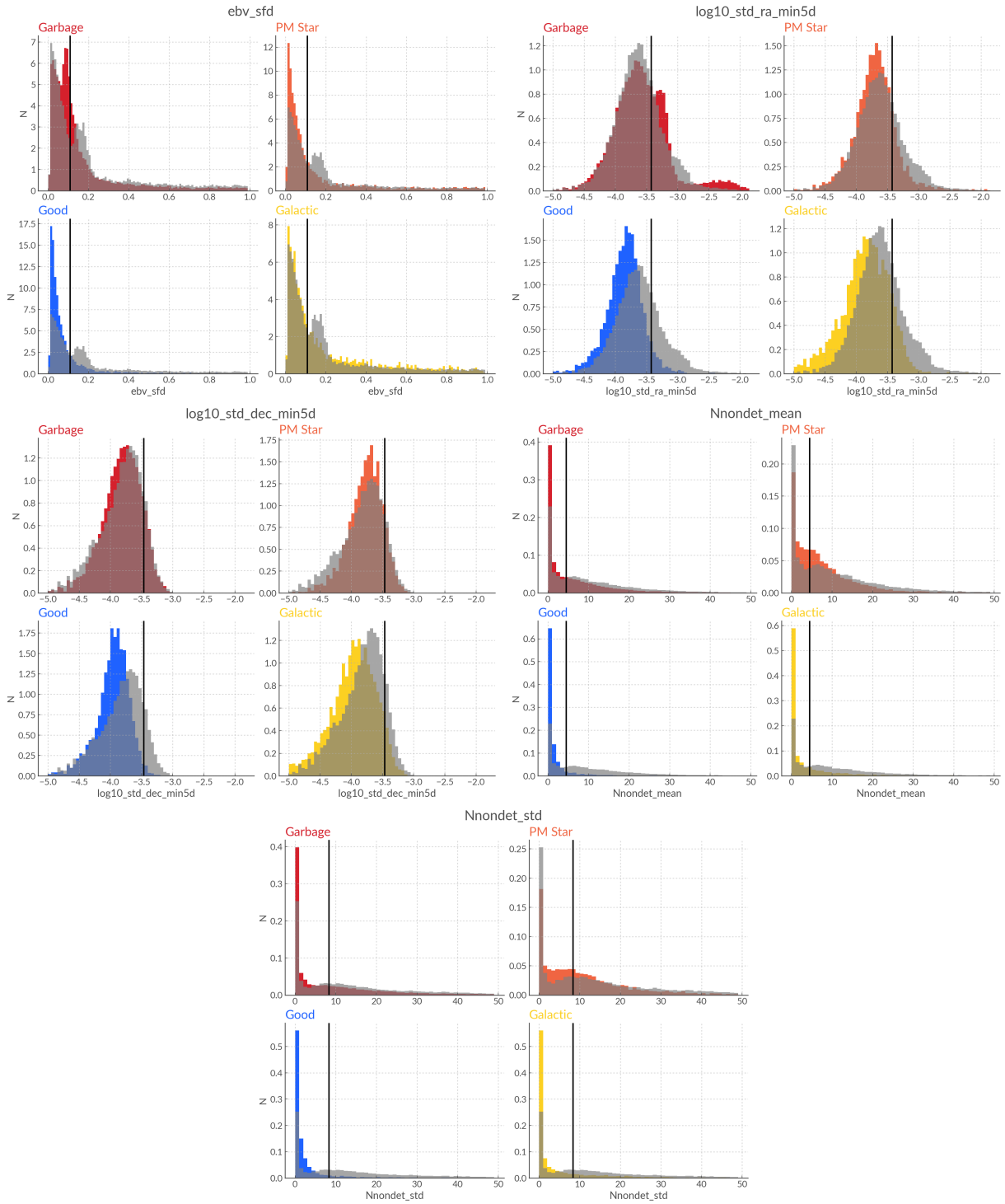

Figure 25. The five day 1 features of AT 2024lwd whose values are anomalous for a "good" object, shown on top of the distributions for these features split by alert type. The gray distribution is the "auto-garbage" distribution, which is superimposed over the human-vetted labels.

Appendix B
On Features and Metrics, Bias and Expertise

As neural networks and transformers have gained popularity, a commonly used justification for choosing these methods is that they do not require feature extraction. The details of why this is an advantage are most often not discussed and it has become common to view not needing to extract or engineer features as an automatic advantage.

A first argument for the use of "feature-free" models is that they allow us to remove a step (or steps) of data processing that can be computationally expensive. What is then omitted however is that a new form of processing is required to make astronomical data usable to neural networks or transformers. To cite a recent example, D. Moreno-Cartagena et al. (2025) studied how visual transformers can be used for photometric classification, stating in their aims that these methods could "classify photometric lightcurves without the need for feature extraction or multi-band

preprocessing." On further inspection the use of visual transformers requires some significant preprocessing steps such as turning multiband lightcurves into images; this was done via matplotlib in their implementation. Lightcurve feature extraction as done in, for example, the VRA or in BTSbot (N. Rehemtulla et al. 2024) is much more lightweight than the generation of images of lightcurves to be parsed to a visual transformer, especially in the context of a large data stream such as LSST. Feature-based models can therefore have a lesser data processing burden than "feature-free" methods, whether they be bespoke neural networks or pretrained large-scale models.[18]

Another argument encountered is that feature extraction is a biased form of preprocessing, or that an algorithm that handles raw data "extracts by itself the best feature representation for a given problem" (J. Pasquet et al. 2019). In a context where it would be untractable for a human to extract meaningful features that can be interpreted by an algorithm, either due to the level of abstraction required or the large quantity and diversity of data (e.g., computer vision), this statement holds true. But sometimes the underlying sentiment is that "raw" data is unbiased (or less biased), and that processing performed by a human inevitably taints these data with unwanted bias. We will take it as granted that the reader agrees that no data is unbiased, and focus on the discussion surrounding human intervention in data processing and to what degree "human bias" is an issue.

Here we understand the term bias to mean a representation of data that is unrepresentative of the characteristics of interest. There are three key areas where bias can be introduced: data collection, data abstraction/preprocessing, and metric choice/interpretation.

An example of biased data collection can be found for example in a kilonova transient classifier reported to have 95% precision (R. Liang et al. 2023) but whose training set contains "contaminant" transients (Types Ia, Ib/c, and II and SLSNe-I), which omit the main contaminants we can expect in a real-life setting (shock breakout of CCSNe and CVs). An example of biased data abstraction is the use of inadequate statistics, such as using the mean of a skewed distribution, which would be biased by the tails (a common everyday example is household income: in 2022 the mean UK income was 39,328 whereas the median was 32,349).[19] Finally an example of poor metric choice can be taken from a medical imaging methodology paper where the authors showed that CNNs trained to detect tumors were biased against finding small tumors (the ultimate goal being early detection) because their performance metric was based on the number of cancerous pixels detected in each image (A. Reinke et al. 2024).

Removing bias can therefore not be achieved by removing a single step of the development process. Feature engineering is a form of data abstraction; even if we assumed that delegating all the data abstraction to the algorithm removes bias from this step (it does not), bias can still be an issue at the data collection level and when creating and evaluating metrics. Bias cannot be eliminated but it can be mitigated and disclosed.

We will take this discussion further and posit that wherever bias can be introduced expertise or domain knowledge can as well. AI for Science professionals can and should take advantage of this.

At the data collection stage, domain knowledge is required to choose training sets that span the full set of characteristics we expect our models to encounter in production. A training set having different properties from a production data set is sometimes referred to as "data drift." This can occur when live data properties slowly change over time from the initial training, but a discrepancy between training and live data can occur as soon as a model is put into the world if the choice of training data is not informed by expert knowledge of the real-life setting.

At the data abstraction stage, expertise can be imbued in the models through feature engineering and feature choice. Even when using methods that can be feature-free, such as CNNs or RNNs, recent successful examples of automation (N. Rehemtulla et al. 2024; X. Sheng et al. 2024; R. Gupta et al. 2025) make use of features or metadata to provide more reliable results. These additional features provide context to the images that researchers know are relevant because of their domain knowledge. It is also worth noting once again that these features can be computationally inexpensive, such as the maximum magnitude and date at maximum computed by BTSBot and the VRA, or the lightning bolt method recently proposed to capture the shock breakout peak and main peak of some CCSNe (A. Crawford et al. 2025). The sole use of raw data when other information is available should be well justified, as it is not obvious that feature-free models are faster and less biased.

Domain knowledge is also essential in designing and choosing metrics that capture the science problems we are addressing. In the case of the VRA, which is designed to help rank eyeball lists to look for extragalactic transients, we assess model performance by measuring the R@K, where recall focuses on extragalactic transients in our sample. This provides a metric that is directly informative, compared to an accuracy, recall, or precision score, which would tell us very little about the value of the ranking.[20]

Finally another area where expertise is essential is in benchmarking and assessing how models compare to state-of-the-art solutions, including non-ML ones. This benchmarking step is not systematically shown in the astro-ML literature, or the comparisons are limited to other ML proofs of concept rather than the currently used methods. Currently it is commonplace for models to be deemed "smarter" when they are more complex (e.g., A. Crawford et al. 2025), which we suspect is the result of the way new advances in ML and artificial intelligence (AI) have been communicated to the general public[21] in the last few years. The reason we would urge science professionals to stay away from these terms is that "smart" and "intelligent" are not descriptive and they are not value-neutral. "Smarter" is better than "less smart." Yet, more complex models are not necessarily better. They have the potential to capture more complex data abstractions, which may be completely irrelevant to a given use case, leading to the creation of bots that, at best, are unnecessarily complex to understand for future team members or users and unnecessarily difficult to maintain; at worst, they generalize poorly (overfit) and introduce uninterpretable layers of data processing that

[18] Sometimes referred to as "foundational models."

[19] https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/householddisposableincomeandinequality/financialyearending2022

[20] It also makes the models biased toward the fast recovery of extragalactic transients, at the expense of Galactic transients, but that is a design choice that is known and reported.

[21] And since astronomers are not primarily AI professionals we are the general public in that context.

future generations of scientists will have to wrestle with. Additionally, different ML methods are developed and specialized for different uses; choosing a model that is best suited to one's given type (and volume) of data is preferable to choosing a model that is more complex but built for a different data type, or used in settings where the amount of data available for training is far superior to the amount of data available in one's field.

Overall, we hope to remind our colleagues that larger ML models and larger data sets rarely mitigate the effects of unrepresentative data, poor design, and irrelevant metrics; we refer the reader to D. Huppenkothen et al. (2023) for an extended discussion of ML best practices in astronomy.

ORCID iDs

H. F. Stevance https://orcid.org/0000-0002-0504-4323
K. W. Smith https://orcid.org/0000-0001-9535-3199
S. J. Smartt https://orcid.org/0000-0002-8229-1731
N. Erasmus https://orcid.org/0000-0002-9986-3898
D. R. Young https://orcid.org/0000-0002-1229-2499
A. Clocchiatti https://orcid.org/0000-0003-3068-4258

References

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. 2018, Journal of Choice Modelling, 28, 167
Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002
Buitinck, L., Louppe, G., Blondel, M., et al. 2013, arXiv:1309.0238
Cenko, S. B., Kulkarni, S. R., Horesh, A., et al. 2013, ApJ, 769, 130
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv:1612.05560
Crawford, A., Pritchard, T. A., Modjaz, M., et al. 2025, ApJ, 989, 192
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2019, MNRAS, 484, 93
Dyer, M. J., Ackley, K., Jiménez-Ibarra, F., et al. 2024, Proc. SPIE, 13094, 130941X
Erasmus, N., Potter, S. B., van Gend, C. H. D. R., et al. 2024a, Proc. SPIE, 13096, 130968K
Erasmus, N., Steele, I. A., Piascik, A. S., et al. 2024b, JATIS, 10, 025005
Friedman, J. H. 2001, AnSta, 29, 1189
Gagliano, A., Contardo, G., Foreman-Mackey, D., Malz, A. I., & Aleo, P. D. 2023, ApJ, 954, 6
Gal-Yam, A. 2019, ARA&A, 57, 305
Gal-Yam, A. 2021, AAS Meeting, 237, 423.05
Gezari, S. 2021, ARA&A, 59, 21
Gillanders, J. H., Rhodes, L., Srivastav, S., et al. 2024, ApJL, 969, L14
Green, G. 2018, JOSS, 3, 695
Groot, P. J., Bloemen, S., Vreeswijk, P. M., et al. 2024, PASP, 136, 115003
Gupta, R., Muthukrishna, D., Rehemtulla, N., & Shah, V. 2025, MNRAS, 542, L132
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Heinze, A. N., Denneau, L., Tonry, J. L., et al. 2021, PSJ, 2, 12
Hunter, J. D. 2007, CSE, 9, 90
Huppenkothen, D., Ntampaka, M., Ho, M., et al. 2023, arXiv:2310.12528
Kaiser, N., Aussel, H., Burke, B. E., et al. 2002, Proc. SPIE, 4836, 154
Ke, G., Meng, Q., Finley, T., et al. 2017, in Advances in Neural Information Processing Systems 30 (NIPS 2017), ed. I. Guyon et al. (Red Hook, NY: Curran Associates, Inc.) https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
Killestein, T. L., Kelsey, L., Wickens, E., et al. 2024, MNRAS, 533, 2113
Killestein, T. L., Lyman, J., Steeghs, D., et al. 2021, MNRAS, 503, 4838
Law, N. M., Kulkarni, S. R., Dekany, R. G., et al. 2009, PASP, 121, 1395
Liang, R., Liu, Z., Lei, L., & Zhao, W. 2023, Univ, 10, 10
McElfresh, D., Khandagale, S., Valverde, J., et al. 2023, arXiv:2305.02997
McKinney, W. 2010, in Proc. 9th Python in Science Conf., ed. S. van der Walt & J. Millman (Austin, TX: SciPy), 56
Möller, A., Peloton, J., Ishida, E. E. O., et al. 2021, MNRAS, 501, 3272
Moreno-Cartagena, D., Protopapas, P., Cabrera-Vives, G., et al. 2025, arXiv:2502.20479
Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, PASP, 131, 118002
The pandas development team 2020, pandas-dev/pandas: Pandas, latest, Zenodo, doi:10.5281/zenodo.3509134
Pasquet, J., Pasquet, J., Chaumont, M., & Fouchez, D. 2019, A&A, 627, A21
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Perley, D. A., Fremling, C., Sollerman, J., et al. 2020, ApJ, 904, 35
Perley, D. A., Ho, A. Y. Q., Fausnaugh, M., et al. 2025, MNRAS, 537, 2362
The PLAsTiCC team, Allam, T., Jr., Bahmanyar, A., et al. 2018, arXiv:1810.00001
Potter, S. B., Erasmus, N., van Gend, C. H. D. R., et al. 2024, Proc. SPIE, 13098, 130980Y
Prentice, S. J., Maguire, K., Smartt, S. J., et al. 2018, ApJL, 865, L3
Radhakrishnan Santhakumari, K. K., Battaini, F., Di Filippo, S., et al. 2024, arXiv:2407.17288
Rehemtulla, N., Miller, A. A., Jegou Du Laz, T., et al. 2024, ApJ, 972, 7
Reinke, A., Tizabi, M. D., Baumgartner, M., et al. 2024, NatMe, 21, 182
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Shappee, B. J., Prieto, J. L., Grupe, D., et al. 2014, ApJ, 788, 48
Sheng, X., Nicholl, M., Smith, K. W., et al. 2024, MNRAS, 531, 2474
Smartt, S. J., Clark, P., Smith, K. W., et al. 2018, ATel, 11727, 1
Smartt, S. J., Valenti, S., Fraser, M., et al. 2015, A&A, 579, A40
Smith, K. W., Smartt, S. J., Young, D. R., et al. 2020, PASP, 132, 085002
Stalder, B., Tonry, J., Smartt, S. J., et al. 2017, ApJ, 850, 149
Steeghs, D., Galloway, D. K., Ackley, K., et al. 2022, MNRAS, 511, 2405
Steele, I. A., Smith, R. J., Rees, P. C., et al. 2004, Proc. SPIE, 5489, 679
Stevance, H. 2025a, ATLAS VRA Technical Manual v3, Zenodo, doi:10.5281/zenodo.14944208
Stevance, H. 2025b, ATLAS VRA v1 - Training Data and Code, Zenodo, doi:10.5281/zenodo.15195392
Stevance, H., & Smith, K. 2025, HeloiseS/atlasvras: VRA v1.1, Zenodo, doi:10.5281/zenodo.14983116
Stevance, H. F., Leland, J., & Smith, K. W. 2025, arXiv:2506.06403
Tonry, J. L., Denneau, L., Heinze, A. N., et al. 2018, PASP, 130, 064505
Vernet, J., Dekker, H., D'Odorico, S., et al. 2011, A&A, 536, A105
Weston, J. G., Smith, K. W., Smartt, S. J., Tonry, J. L., & Stevance, H. F. 2024, RASTI, 3, 385
Williams, R. D., Francis, G. P., Lawrence, A., et al. 2024, RASTI, 3, 362
Worters, H. L., O'Connor, J. E., Carter, D. B., et al. 2016, Proc. SPIE, 9908, 99083Y
Young, D. R. 2023, Sherlock. Contextual Classification of Astronomical Transient Sources v2.2.0, Zenodo, doi:10.5281/zenodo.8038057