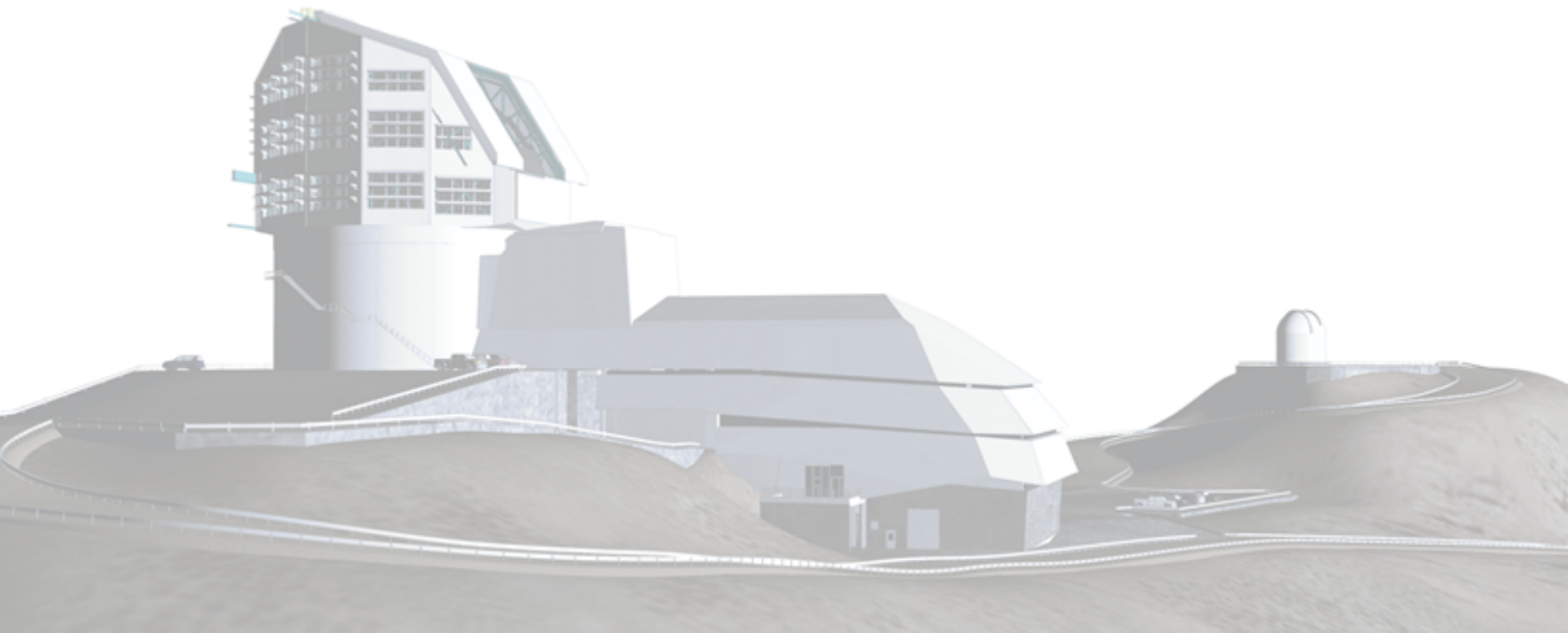
Vera C. Rubin Observatory
Data Management

Performance of Machine-Learned Reliability Scoring for Image Differencing

Tatiana Acero-Cuellar, Federica B. Bianco, Bruno Sanchez, Masao Sako, Eric C. Bellm

DMTN-337

Latest Revision: 2026-05-20



Abstract

This document summarizes the status of the Machine Learning Reliability model (also known as "Real/Bogus") for the LSST alerts stream. The model associates to each Difference Image Analysis (DIA) detection a `reliability` $[0-1]$ score designed to measure the likelihood of the veracity of the detection, *i.e.*, astrophysical source (1) *vs* artifact (0). This manuscript documents the architecture and performance of each model version, including specific performance characteristics, identified weaknesses and failure modes, mitigation strategies, and details of the training set.

Change Record

Version	Date	Description	Owner name
1	2026-05-07	Initial Release.	Acero-Cuellar

Document source location: `https://github.com/lsst-dm/dmtn-337`

Contents

Performance of Machine-Learned Reliability Scoring for Image Differencing

1 Introduction

Rubin requirements from *Claver & The LSST Systems Engineering Integrated Project Team (LSE-30) define the Difference Source Spuriousness Threshold - Transients | ID: OSS-REQ-0353* stating the following: "There shall exist a spuriousness threshold T for which the completeness and purity of selected difference sources are higher than `transCompletenessMin` and `transPurityMin`, respectively, at the SNR detection threshold `transSampleSNR`. This requirement is to be interpreted as an average over the entire survey." For Transients, the thresholds are:

- `transCompletenessMin`: 90%,
- `transPurityMin`: 95%,
- `transSampleSNR`: 6.

Additionally, from Graham et al. (DMTN-102) "It is a requirement that the Data Management System be capable of supporting the distribution of at least 98% of alerts for each visit within 60 seconds of the end of image readout.".

To help meet these requirements, we developed a light Convolutional Neural Network that takes as input the postage stamps: template, science, and difference, as they are produced by the Difference Image Analysis (DIA) Pipeline, and returns a reliability score, `reliability`, (a.k.a, `Real`/`Bogus` score), as a number between 0 and 1. This is a supervised machine learning model, a binary classifier where the labels for each triplet of postage stamps are either `Real` or `Bogus`. The labels are gathered with a combination of fake injection, and labeled detections obtained by analyzing classifications given by volunteers, both Rubin Observatory members and citizen scientists participating in the Zooniverse[1] project called 'Rubin Difference Detectives'.[2] This note reports the performance of that classifier, the 'Rubin Machine Learning (ML)-reliability model'.

[1] `https://www.zooniverse.org/`.
[2] `https://www.zooniverse.org/projects/ebellm/rubin-difference-detectives`.

The set of model weights that generates the reliability score has been updated three times as of April 2026. While the model architecture remains the same, the weights are produced by training, retraining, or fine-tuning on different datasets. Each version of the reliability model is detailed in Table 1, Table 2.[3]

Model version number	Butler collection name	Training inputs
0.1	tac_cnn_comcam_2025-02-18	DC2 + LSSTComCam + injections
0.2	tac_cnn_lsstcam_2026-02-13	(DC2 + LSSTComCam + injections) + zooniverse
0.3	tac_cnn_lsstcam_2026-02-26	(DC2 + LSSTComCam + injections) + zooniverse + LSSTComCam + injections

TABLE 1: Model versions. Model `v0.1` represents the base model, version `v0.2` and `v0.3` were obtained by fine-tuning the base model. The training inputs in parentheses are data used to train the base model, but were not directly used to fine-tune the respective model.

Model version number	AP deployment date	AP reliability cutoff	DRP usage
0.1	September 2025	0.1	DP1 model
0.2	18 February 2026	0.1 (before Feb 24, 2026) 0.5 (after Feb 24, 2026)	
0.3		0.5	DP2 model

TABLE 2: Model version cont.ed. The AP reliability cutoff is implemented to mitigate the impact on APDB when generating alerts. The AP and DRP pipelines worked independently; each one can use a different model version. `v0.2` was used for the initial alerts, and as of April 2026, it is still the model used for the alerts.

2 Description of the Machine Learning-reliability model

The following section describes the Convolutional Neural Network architecture used to train on the postage stamps and predict the reliability scores for each Difference Image Analysis (DIA) detection. We developed a relatively simple model: a Convolutional Neural Network with three convolutional layers and two fully connected layers. The convolutional layers have a 5×5 kernel size, with 16, 32, and 64 filters, respectively. A max-pooling layer of size 2 is applied at the end of each convolutional layer, followed by a dropout layer of 0.4 to reduce overfitting. The last fully connected layers have sizes of 32 and 1. The ReLU activation function is used for

[3] By 'fine-tuning' here we refer to the practice of retraining a model starting training from the weights set in previous models.

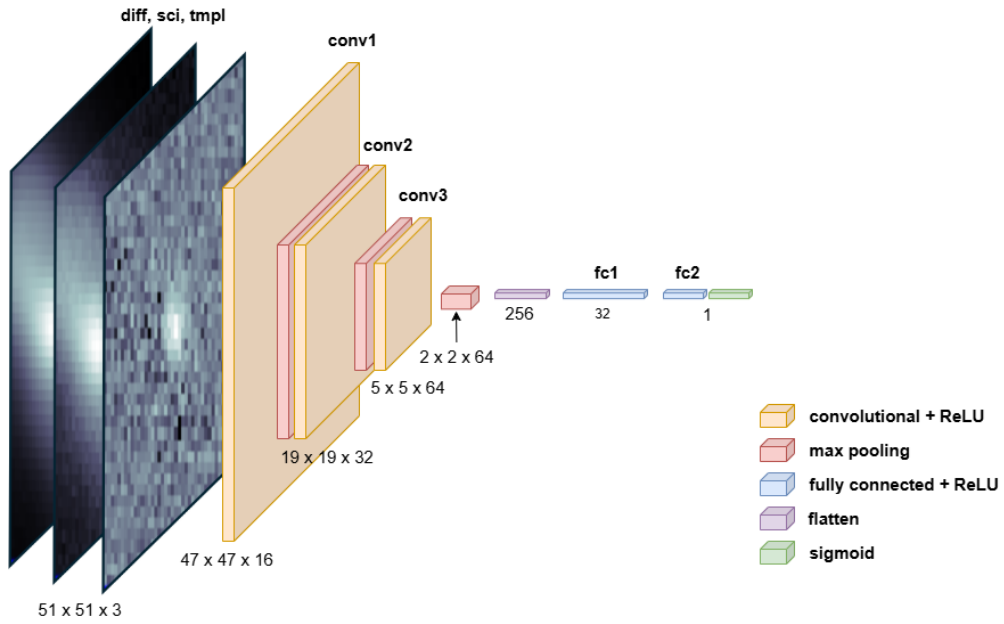


FIGURE 1: CNN architecture of the Machine Learning model that generates a reliability score for every DIASource. The input data are the combined template, science, and the difference image postage stamps, each of size 51×51 pixels, centered on the detected sources. The input array has a shape of (3, 51, 51). The CNN has three convolutional layers and two fully connected layers. The last layer has an output size of 1, and the sigmoid function is used for the output layer to provide a probabilistic interpretation. The design of the CNN is purposefully minimal to meet computational and time requirements.

the convolutional layers and the first fully connected layer, while a sigmoid function is used for the output layer to provide a probabilistic interpretation. The cutouts are generated by extracting postage stamps of 51×51 pixels centered on the detected sources. The input data of the model consists of the template, science, and difference image stacked to have an array of shape (3, 51, 51) (See Figure 1). The model is implemented using PyTorch (Paszke et al., 2019). The Binary Cross Entropy loss function is used, along with the Adaptive Moment Estimation (Adam) optimizer with a fixed learning rate of 1×10^{-4}, weight decay of 3.6×10^{-2}, and a batch size of 128. The design of the CNN is guided by the need to meet computational and timing requirements as specified in the 'Data Management System Requirements' (DMSR, Dubois-Felsmann & Jenness (LSE-61)). While the model is architecturally conservative, we continuously test its performance against that of other models of demonstrated accuracy (e.g., Inada et al., 2026), so as to direct our improvement towards either the model or the training data. At the time of writing, we identified that the most significant improvements can be achieved by improving the training data.

3 Data

The data to train the model has been in constant evolution, adjusting it every time to the most recent status of the telescope and data pipelines. For each model update, a specific dataset was used for training. Here, we described the three datasets used for `v0.1`-`v0.3` models.

3.1 Data for `v0.1` (DP1) model: LSSTComCam data

The ML-Reliability model was initially trained with 89,066 `Real` and the same amount of `Bogus` labels, extracted from the Data Challenge 2 simulations (DC2, LSST Dark Energy Science Collaboration et al., 2021), plus random injections of stars (i.e., PSFs at random locations without Galaxy associations) to increase the number of `Real`. Once the LSSTComCam data became available, the model was fine-tuned on a subset of that data containing 183,046 postage stamps with PSF injections. This fine-tuned model was the one used to generate reliability scores for Data Preview 1 (DP1, Vera C. Rubin Observatory Team, RTN-095) and is tagged `v0.1`.

Name	handle	# sources	Pipeline	Type
tns_ap	tns	219	AP	TNS match
tns_drp33	tns	855	DRP 2025_33	TNS match
tns_drp37	tns	984	DRP 2025_37	TNS match
galaxy_ddf	galaxy	10,932	AP	Galaxy match (extendedness=1)
rel_cuts_drp33_gt_05_lt_09	reliability_cuts	49,933	DRP 2025_33	reliability cut
ss_drp37	ss	88,058	DRP 2025_37	Solar System match
gaia_drp37	gaia	89,410	DRP 2025_37	GAIA match

TABLE 3: Description of the DIA sources extracted from AP and DRP data by cross-matching with GAIA and TNS catalog, with the internal Solar System Catalogs, and with the sources with parameter `extendedness=1` given by Rubin source catalogs, and transients that received a $0.5 < \texttt{reliability} < 0.9$ score from the `v0.1` model. 'Handle' refers to the short-hand label by which these sets will be referred to throughout the rest of this manuscript.

3.2 Data for `v0.2` and `v0.3` (DP2) model: LSSTCam and Zooniverse

The model was retrained (fine-tuned) to improve performance on the commissioning LSSTCam data that will eventually be released as Data Preview 2 (DP2 AlSayyad & O'Mullane, RTN-111), leading to model version `v0.3`. The `v0.2` model, as of April 2026, is being used for the AP pipeline and the production of alerts. The training of `v0.3`, and `v0.2` relies partially and totally, respectively, on real data (no fake injection) and `Real`/`Bogus` labels obtained by the Zooniverse citizen scientists through the 'Rubin Difference Detectives' project. The fine-tuned model will be used to generate reliability scores for DP2.

3.2.1 Rubin Difference Detectives: Zooniverse Citizen Science Project

Rubin Difference Detectives was launched on November 11th, 2025, with 242,391 DIA Sources (a "subject" in Zooniverse terms) for the Zooniverse citizen scientists to label. The distribution of the DIA sources per catalog is described in Table 3. By associating DIA detections in LSSTCam data with various catalogs (TNS (Transient Name Server), Gaia (Rimoldini et al., 2023), Rubin Solar System catalogs), and DIA detections with `reliability` score between 0.5 and 0.9, `Real` labels generated by the `v0.1` model (subset tagged 'reliability_cuts')), and by using measured DIA source properties, we produced a dataset of DIA detections that is expected to be close to balanced in the `Real`/`Bogus` distribution.

The subjects given to the citizen scientists were images composed of three postage stamps of size 51×51: template, science, and difference images, the DIA pixel output for each detected source. Volunteers were asked to perform the following task: *Label the object seen in the center of the difference image as* `Real` *or* `Bogus`; no skip option was provided. Figure 2 shows how the

image, the task, and the answer options were presented to the volunteers.

FIGURE 2: Zooniverse interface of the Rubin Difference Detectives project. The citizen scientists were shown a subject, which is an image with the template, science, and the difference postage stamps, and they were asked to select whether the source in the difference image is `Real` or `Bogus`.

For a subject to be "retired", meaning volunteers can no longer see it and labels are no longer collected, it first had to have been classified as either `Real` or `Bogus` by two different volunteers in agreement. If this initial condition was not met, the subject was instead retired after having been classified by seven different volunteers.

The following section describes how the classifications given to a DIA source were used to determine the final `Real` or `Bogus` label, which, in the end, corresponded to the ground truth labels to train the reliability score model.

3.2.2 Analysis of the Zooniverse classifications

The methodology we used to understand the classifications made by the volunteers and estimate the `Real`/`Bogus` classifier performance of the Rubin Difference Detectives project was a modification of the algorithm defined and implemented in Marshall et al. (2016). Following this work, we implemented a probabilistic classification for each subject (Pr(`Real`): probability of the DIA source to be a `Real` astrophysical object). Their probabilistic methodology relied on understanding how the volunteers responded to a reference data set (subjects where the ground truth was known) and thus building a prior for each user. The probabilistic classifier is explained in great detail in Marshall et al. (2016). We refer the reader to their paper. Our

reference data set was composed of either the DIA sources classified by the Oxford team (8 experts from the University of Oxford who classified a fraction of the DIA sources in Zooniverse (Weston et al. in preparation)), the DIA sources classified by the Rubin-ML-Reliability team (referred to as Rubin hereafter), or both. The reference set had a total of 20,261 (`Real`: 8576, `Bogus`: 11,686) DIA sources, when considering both reference datasets, and a total of 8956 (`Real`: 3805, `Bogus`: 5151) DIA sources, when considering only the Rubin expert labels. We note here that we only consider volunteers who provided labels for at least one subject in the reference set. At the time of training model version `v0.2`, out of 3750 volunteers, only 2476 encountered a DIA source of the Rubin reference data set (2792 when considering both reference datasets).

Because the number of labels is typically much smaller in 'Rubin Difference Detectives' compared to the 'Space Warp' project that inspired Marshall et al. (2016), we implement the following modifications:

- We set a minimum threshold of 0.05 and a maximum threshold of 0.95 to the participants' Bayesian prior to avoid collapse of the posterior (e.g., if a volunteer only labeled one reference source, their prior would be 0 or 1, depending, causing the posterior to collapse to the same value);
- We implement a stability criterion to accept a subject's label as described below.

Since the Pr(`Real`) per DIA source returned a float between 0 and 1, to determine the final binary classification (`Real` = 1 or `Bogus` = 0), we defined two different scenarios.

1. *stable set*: A source is considered stable if (1) the probability of being `Real`, Pr(`Real`), meets a threshold (≥ 0.90 for `Real`, ≤ 0.10 for `Bogus`) and (2) this condition holds for at least the last three classifications.
2. *unstable set*: Only the threshold condition applies, without the stability criterion.

In Figure 3, we showed the fraction of Positives (P) and Negatives (N) labels ($N_{\mathrm{P}}/N_{\mathrm{P\ labels}}$ and $N_{\mathrm{N}}/N_{\mathrm{N\ labels}}$ respectively) obtained by comparing volunteer classification and reference data labels as a function of the number of labels provided by the volunteer. In general, the more subjects a volunteer labels, the higher the P and N fractions. Ideally, the P and N fractions

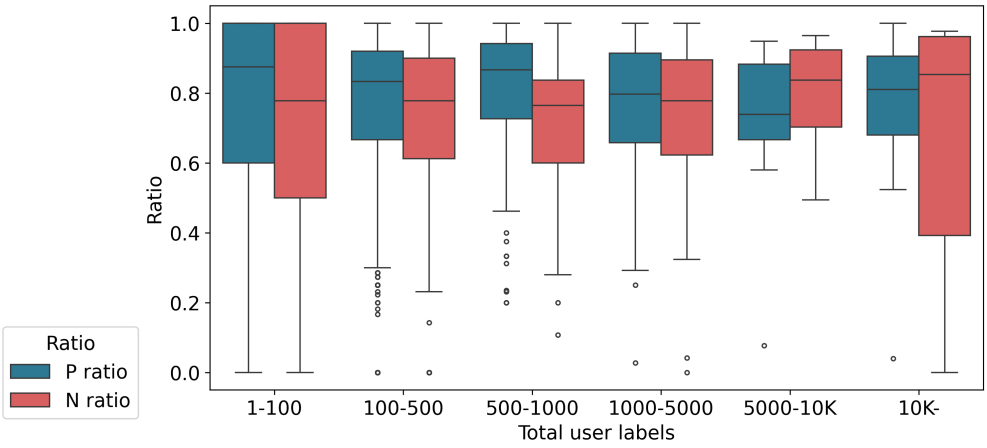


FIGURE 3: Fraction of Positives (P) and Negatives (N) labels obtained by comparing Zooniverse volunteer classification and reference classification (perfect agreement with the reference label would lead to a ratio of 1), per total classifications made by the user. The larger distributions for the 1-100 and >10k user labels is due to small number statistics. There are some outliers where the volunteers are very optimistic and classify everything as `Real`, or very pessimistic and classify everything as `Bogus`. Those users could be discarded, but in the Bayesian framework described in Marshall et al. (2016), the impact of their labels is naturally suppressed.

should be similar, indicating that the user can distinguish `Real` and `Bogus` sources and is not biased toward one class.

The distribution of `Real`/`Bogus` for the reference data changes depending on the type of source (`gaia`, `galaxy`, `reliability_cuts`, `ss`, or `tns`, see Table 3), and it is also related to the Signal to Noise Ratio (SNR) of the DIA source. Visual inspection of some DIA sources labeled by experts (Figure 4) showed that low SNR tends to be classified as `Bogus`, regardless of the type of source.

After applying the Bayesian approach (Marshall et al. 2016) to the Zooniverse classification, the breakdown of labels is described in Table 4.

Scenario	Total `Real`	Total `Bogus`	Reference Data
stable	8236	40,402	Rubin
unstable	25,218	53,024	Rubin
stable	8025	46,124	Oxford and Rubin
unstable	23,889	57,570	Oxford and Rubin

TABLE 4: Final count of `Real` and `Bogus` obtained after determining the probability of the source being `Real` given the classifications provided by the Zooniverse volunteers following the criteria described in subsubsection 3.2.2.

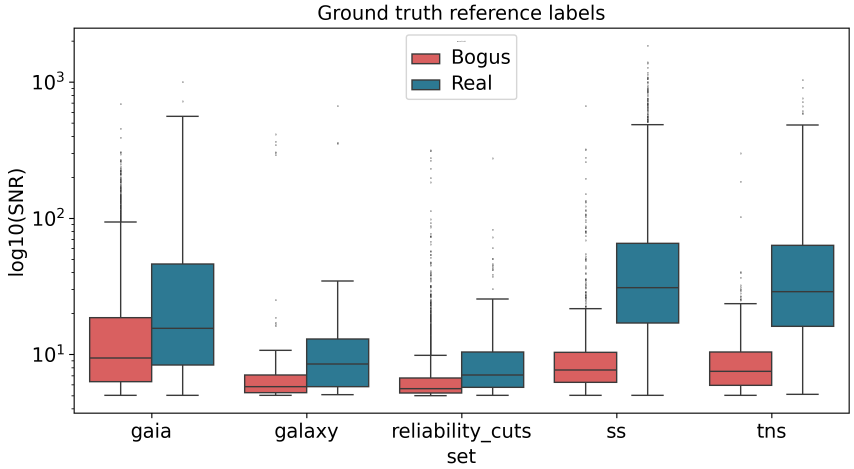


FIGURE 4: Signal to Noise Ratio (SNR) distribution for each source in the respective catalogs for DIA sources labeled by experts as `Real` (blue) or `Bogus` (red).

The total number of `Real` and `Bogus` labels after the classification analysis is highly unbalanced; for the purpose of training a binary classification machine learning model, the data sets should be balanced between classes to avoid inducing a bias. The model performance results shared hereafter are based on a balanced data set, produced by subsampling the data shown in Table 4. The distribution of the `Real` and `Bogus` by data origin, as explained in Table 4, after analyzing the Zooniverse classifications, is shown in Figure 5.

The construction of the ground truth labels for fine-tuning the `v0.2` and `v0.3` models relied on the Zooniverse analysis obtained by only considering the Rubin reference data set (see Table 4). Additionally, only detections with high-confidence labels (stable) were used. In total, 13,178 sources were used to fine-tune the `v0.2` and `v0.3` models: 6,630 `Real` and 6,548 `Bogus`. 1,647 were used to validate, and another 1,647 to test the model. Given the small size of the training data set, in addition to the original images, two augmentations (vertical and horizontal flipping) were added to the training. For `v0.3`, a fraction of the dataset used for the `v0.1` DP1 model (LSSTComCam fakes) was also used. For `v0.2` the fine-tuning relies 100% on the Zooniverse data.

3.3 Data for Future models

Future model iterations will include ground truth labels for fine-tuning the model, relying on the Zooniverse analysis obtained by considering the training data set of both the Oxford team

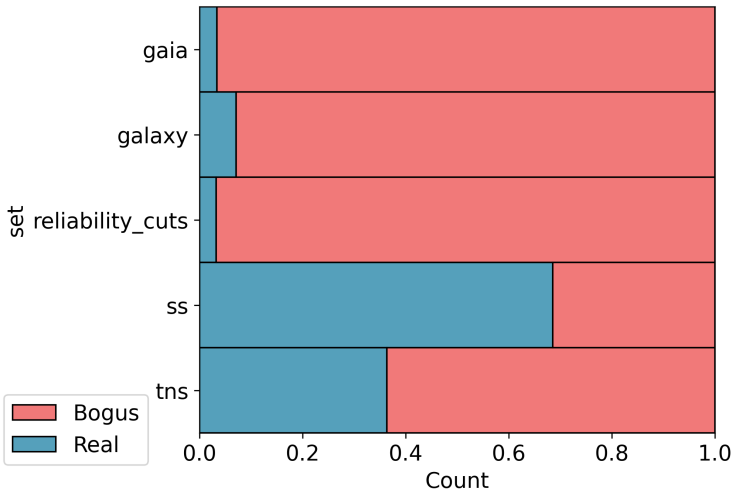


FIGURE 5: Distribution of `Real` and `Bogus` as labeled via the 'Rubin Difference Detectives' Zooniverse project for data used for training/validation/testing of the ML-reliability model `v0.2` and `v0.3`. The data is grouped as in Table 3. Almost 80% of the `Real` labels correspond to Solar System objects ('ss'), 10% correspond to variable objects matched with the GAIA catalog, and the other 10% is distributed between transients (TNS catalog), objects with extendendness = 1 ('galaxy'), and `Real` found by the `v0.1` model ('reliability_cuts'). On the other hand, almost 60% of the `Bogus` objects correspond to matches with the GAIA catalog, which we find are mainly due to bad subtractions, 25% correspond to DP1 false positives ('reliability_cuts'), and the remaining 15% correspond to artifacts in the Solar System objects ('ss'), objects with extendendness = 1 ('galaxy'), and a few artifacts from matches with the TNS catalog.

and the Rubin-ML-Reliability team (see Table 4). Additionally, fake sources injected in more recent injection runs can be included.

4 Performance

The distribution of the `reliability` score for a random set of DIA sources is shown in Figure 6 for all three model versions. Besides the peaks at 0 and 1, each model presents a third peak at ~ 0.5 for `v0.1` model, and ~ 0.3 for both the `v0.3` and `v0.2` models. The peak corresponds to when the science and difference images contain some rows or columns full of `NaN` values.

Given the high volume of DIA sources produced by the AP at the moment of writing, and to mitigate the impact on APDB, when generating alerts and identifying detections for downstream analysis (e.g., forced photometry), a reliability value cutoff is implemented for each model. The cutoff is determined by analyzing the behavior of the model against properties of the DIA sources, and the behavior on fakes, e.g. Figure 7, Figure 8, Figure 9, and Figure 10. This cutoff

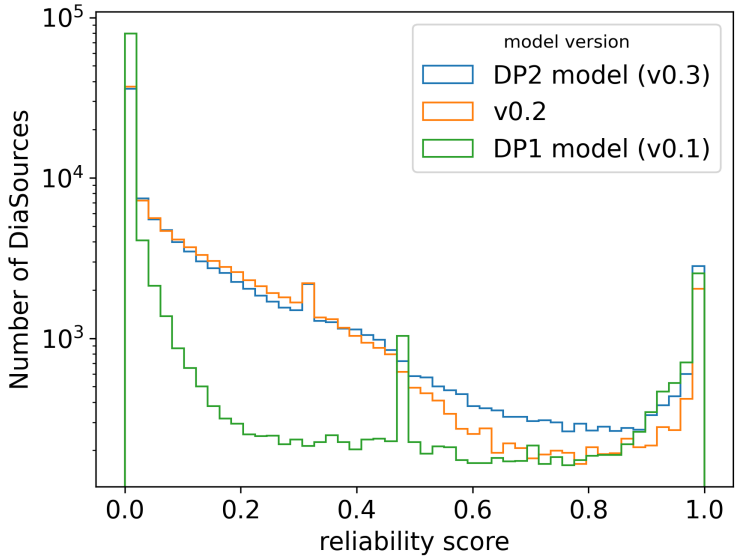


FIGURE 6: Distribution of the reliability score for each model version.

also reduces the amount of Bogus that reaches the Brokers and the scientific community. The current cutoff values are 0.1, 0.5, and 0.5 for `v0.1` model, `v0.2`, and `v0.3` model, respectively. We expect to lower that threshold for future versions of the ML-Reliability model.

The behavior of each model against SNR for a fake injection run outside of the training data set (see subsection 4.1) is shown in Figure 7. As a reminder to the reader, the `v0.1` model was trained only using fakes; `v0.2` was only trained with real data without fake injection, and the `v0.3` model was trained with both fakes and real data. The slope of the distribution of `reliability` for `v0.2` as a function of SNR in the region $0 \leq \text{SNR} \lesssim 12$ indicates a model bias which we attribute to a bias in the volunteer's label, as discussed in subsubsection 3.2.2 (see Figure 4). This bias is mitigated by the inclusion of fakes (injections) in the `v0.3` model data.

Following the basic idea behind the DIA, where the flux of the DIA source F_{Diff} in the difference image is expected to be the subtraction of the science flux from the template flux, we show the flux measured by aperture photometry on the difference image F_{AP} against the flux measured in the science and template images F_{Diff} in Figure 8. In an ideal case, these two values should follow a 1-1 relation (linear relation), and we would expect to have high reliability scores for DIA sources that fall on or near the diagonal in Figure 8. Deviation from that diagonal may indicate some deviation from the basic flux measurement, and therefore a potential Bogus object, which should produce a low reliability score. Figure 9 shows the same data but with the current reliability cutoffs applied (`reliability` $= 0.1$ for `v0.1`, `reliability` $= 0.5$ for `v0.2`

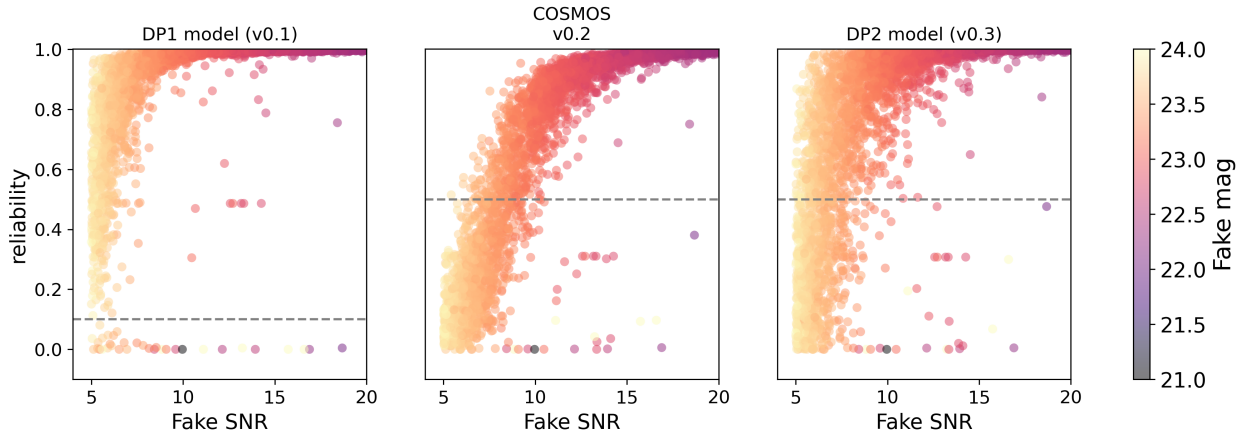


FIGURE 7: Reliability score against SNR for each model version for a set of fakes DIA sources in the magnitude range $m = 24 - 21$ from an injection run outside of the training set. The horizontal dashed line indicates the `reliability` cutoff for each model version: 0.1 for `v0.1` and 0.5 for both `v0.2` and `v0.3`.

and `v0.3`). Generally, we see a strong and encouraging dependence of `reliability` with the ratio of DIA and science-template flux, with deviation mostly near (F_{AP}, F_{Diff}) =(0,0), where low SNR sources tend to get low `reliability` scores, as seen before. We also see an excess of `reliability` > 0.5 for `v0.2` and `v0.3` throughout the lower left quadrant of the plot ($F_{\mathrm{AP}} < 0$ and $F_{\mathrm{Diff}} < 0$), i.e. negative flux sources), while these sources receive extremely low scores in `v0.1` because this model was not trained on negative flux detections. We note that negative sources are likely to be uncharacteristically common in the current DP1 and DP2 DIA due to the proximity in time of the template generation with the science image collection, and to the short timeline during which the templates were constructed.

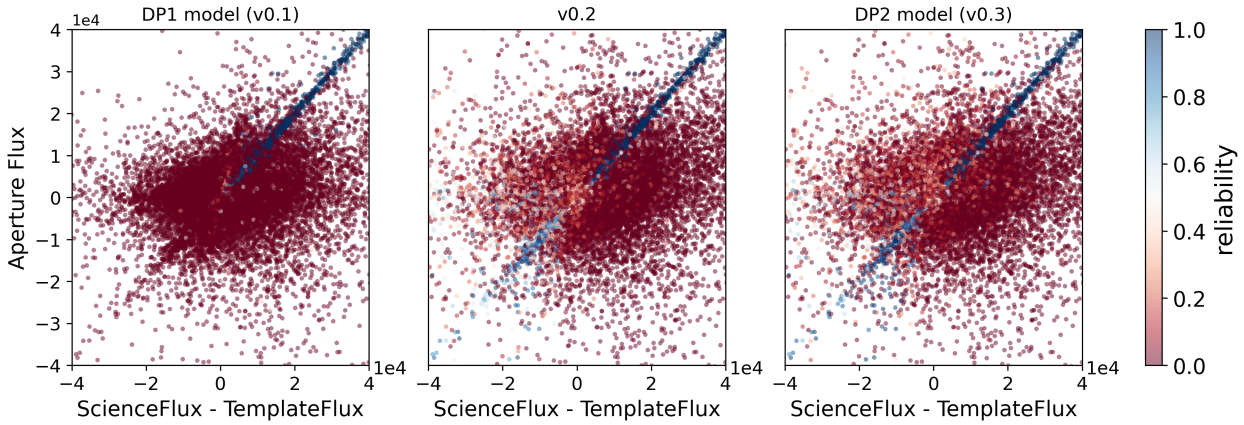


FIGURE 8: DIA Aperture Flux vs ScienceFlux - TemplateFlux colored by the reliability score for each model version for a random set of DIA sources.

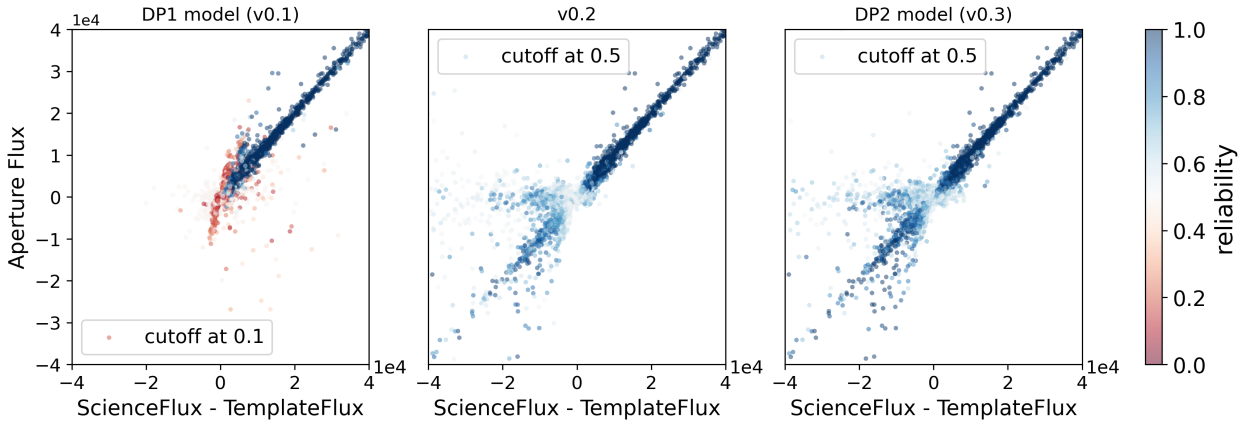


FIGURE 9: DIA Aperture Flux vs ScienceFlux - TemplateFlux colored by the reliability score with a cutoff for each model version for a random set of DIA sources.

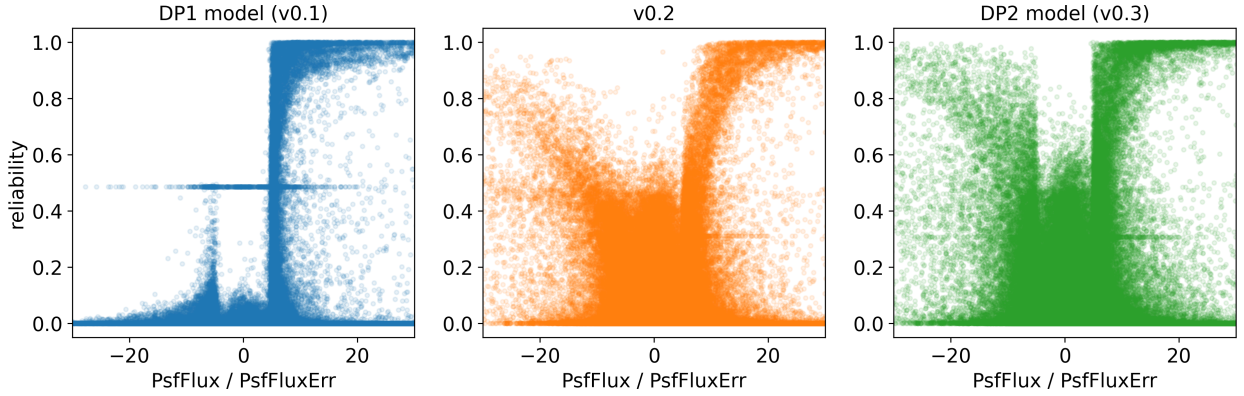


FIGURE 10: `reliability` vs psfFlux/psfFluxErr for each model version for a random set of DIA sources.

The ratio of the quantity `psfFlux` (Flux for Point Source model) and the respective error (`psfFluxErr`), against the reliability score for each model, is shown in Figure 10. In general, DIA sources with large psfFlux error have a reliability score smaller than ~ 0.5. For the `v0.1` model, where DIA sources with negative flux were not considered as part of the training set, the model assigned to all DIA sources with negative flux values a reliability score `reliability` < 0.5. The `reliability` in `v0.3` shows a less obvious functional dependence on the ratio of `psfFlux` to its error; sources in the region $-5 \lesssim$ `psfFlux/psfFluxErr` $\lesssim 5$ are unlitley to get `reliability` > 0.6, but sources with values up to this region result in any value of `reliability`, while in `v0.2` sources with `psfFlux/psfFluxErr` $\lesssim -5$ were less likely to result in high `reliability` scores with a visible direct correlation of `reliability`, with |`psfFlux/psfFluxErr`|. This effect is likely related to the diminished dependence of `reliability` on SNR, as described above.

A quantitative assessment of the performance of each version of the Rubin ML-Reliability model is reported in the following sections.

4.1 Model `v0.1` (deployed on DP1)

Throughout, the performance of the models is measured on all DIA sources, i.e., on any DIA detection with SNR>5 (while the LSST requirements are set to SNR=6).

The `v0.1` model was deployed on DP1 data. The model had achieved a purity of 98.90% and completeness of 97.00% on the test dataset generated from the same collection from which the training and validation data were produced (DC2 simulations, LSSTComCam images with and without injections, see section 3). On a separate LSSTComCam test set, where `Reals` are comprised entirely of injections, the model achieved an accuracy of 98.06%, purity of 97.87%, and completeness of 98.27%, meeting construction requirements.

To further assess the model performance, we used a dataset of fakes injections into the COSMOS field LSSTCam images, which we hereafter refer to as 'COSMOS injections'. This dataset is completely separated and generated independently of the training data set (Table 3). The performance of `v0.1` on this dataset is shown in Figure 11. A purity and completeness of 87% are achieved at a threshold of `reliability`=0.907.

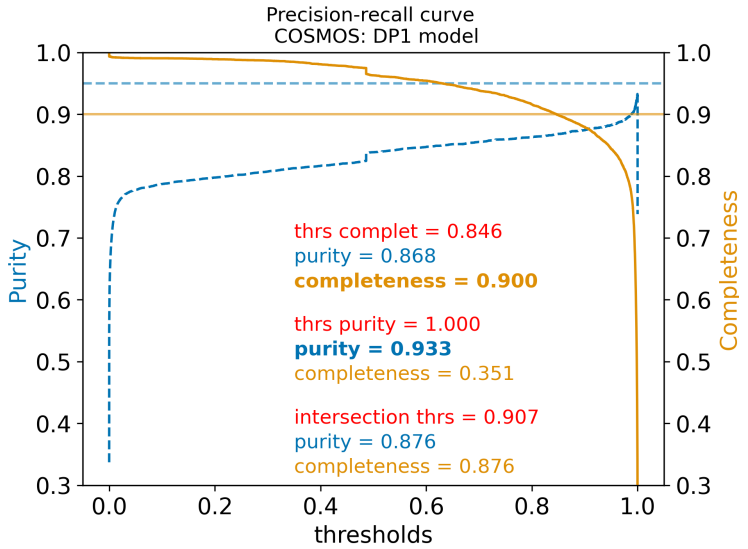


FIGURE 11: Performance of the `v0.1` (DP1) model on fake sources injected into the COSMOS field LSSTCam images.

As reported in (Vera C. Rubin Observatory Team, RTN-095, DP1), and shown in Table 5, when tested on real LSSTComCam detection generated by cross-matching DIA sources as discussed in section 3, and after applying a threshold of `reliability`=0.5, the purity and completeness achieved on Solar System detections is high, on transient detections completeness is $\sim 75\%$, but completeness is very low on variable stars. Since variable stars were not included in the training dataset, and they have specific complexities for DIA in the subtraction of a PSF from an existing PSF (unlike transients and solar system detections), it is to be expected that the model would not perform well on this class. This motivated the development of the Zooniverse project.

TABLE 5: Classification Results for Solar System Transient and Variable objects using the DP1 model with a reliability threshold of 0.5.

Type	True Positives	False Negatives	Total Objects
SS	93.5%	6.5%	5988
Transients	73.7%	26.3%	99
Variables	3.5%	96.5%	316

4.2 Model `v0.2` deployed on initial alerts

The `v0.2` was deployed on the initial LSST alerts (February 2026). The `v0.2` model was tested against the COSMOS injections. Results are shown in Figure 12.

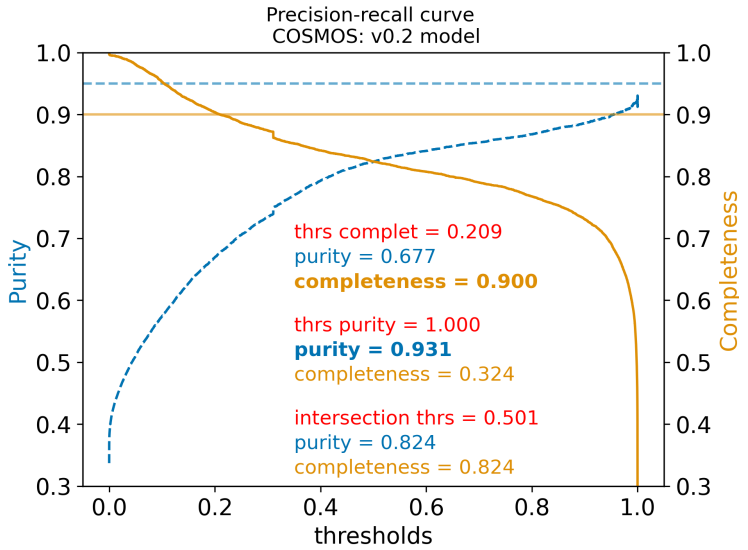


FIGURE 12: Performance of the `v0.2` model on fake sources injected into the COSMOS field LSSTCam images.

The optimal balance of purity and completeness is achieved at `reliability`=0.501 (82% purity

and 82% completeness).

4.3 Model v0.3 (deployed on DP2)

Model v0.3 was benchmarked for processing DP2 data.

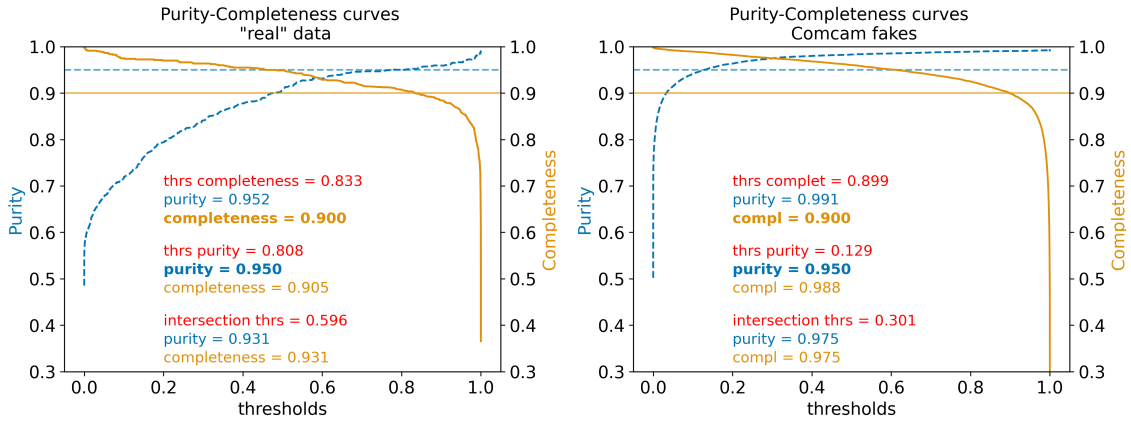


FIGURE 13: Model v0.3 Purity and Completeness curves. *Left*: For the test data without fake injections (Zooniverse labels), at a threshold of 0.596, the purity (precision) and completeness (recall) are both 93.1%. *Right*: For the LSSTComCam fakes test data, at a threshold of 0.301, the purity (precision) and completeness (recall) are both 97.5%. Users can choose their own reliability threshold to trade off completeness vs. purity.

In testing, the model reached a purity $\geq 95\%$ and completeness $\geq 90\%$ with threshold values $0.808 \leq \texttt{reliability} \leq 0.833$ on the test LSSTCam data without fake injections. On the LSSTComCam injections test set, the model reached a purity $\geq 95\%$ and completeness $\geq 90\%$, obtained with values $0.129 \leq \texttt{reliability} \leq 0.899$. Figure 13 shows the corresponding purity and completeness curves.

The v0.3 model performance on the COSMOS injection is shown in Figure 14. 84% Purity and 84% Completeness are achieved at $\texttt{reliability} = 0.785$.

The distribution of v0.3 reliability per data origin is shown in Figure 15. Once again, we see that nearly all Solar System detection receive a $\texttt{reliability} \sim 1$ score while the reliability for other data sources varies, with Gaia and the high confidence v0.1 labels ('reliability_cuts') receiving generally low reliability.

The full Purity and Completeness curves per data origin as measured on the original test data are shown in Figure 16.

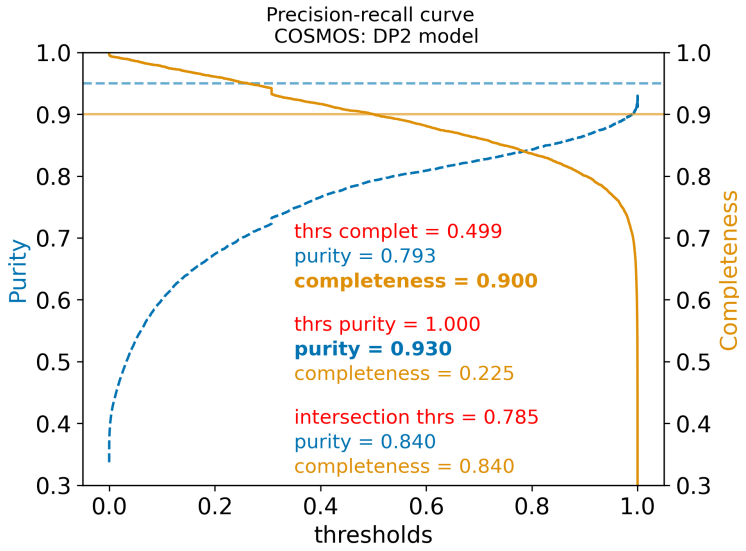


FIGURE 14: Performance of the DP2 model on fake sources injected into the COSMOS field LSSTCam images.

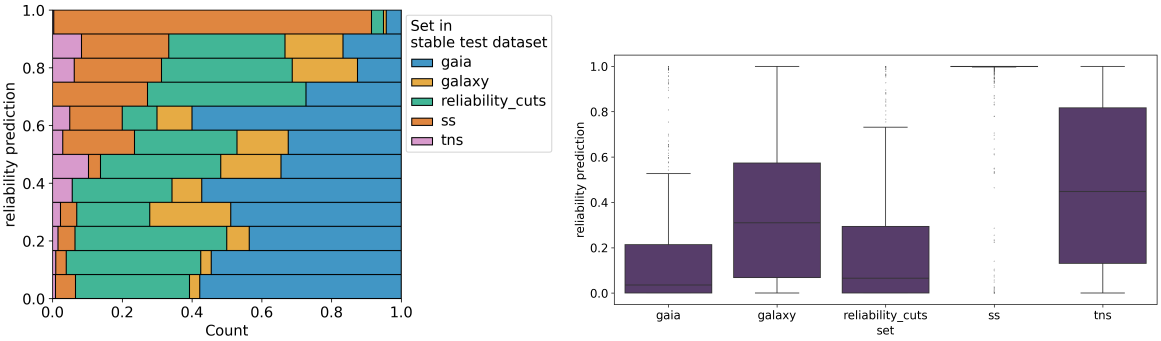


FIGURE 15: Reliability score on test data without fake injections. On the left, the fraction of labels by score range. The majority of high scores come from matches with Solar System objects (see section 3) while all other sets contribute to all score ranges. On the right, while the distribution of Solar System objects is highly compressed near `reliability` = 1 the Gaia cross-matches and high `v0.1` score LSSTComCam detections have characteristically low scores.

A Acknowledgements

This material is based upon work supported in part by the National Science Foundation through Cooperative Agreements AST-1258333 and AST-2241526 and Cooperative Support Agreements AST-1202910 and AST-2211468 managed by the Association of Universities for Research in Astronomy (AURA), and the Department of Energy under Contract No. DE-AC02-76SF00515 with the SLAC National Accelerator Laboratory managed by Stanford University. Additional Rubin

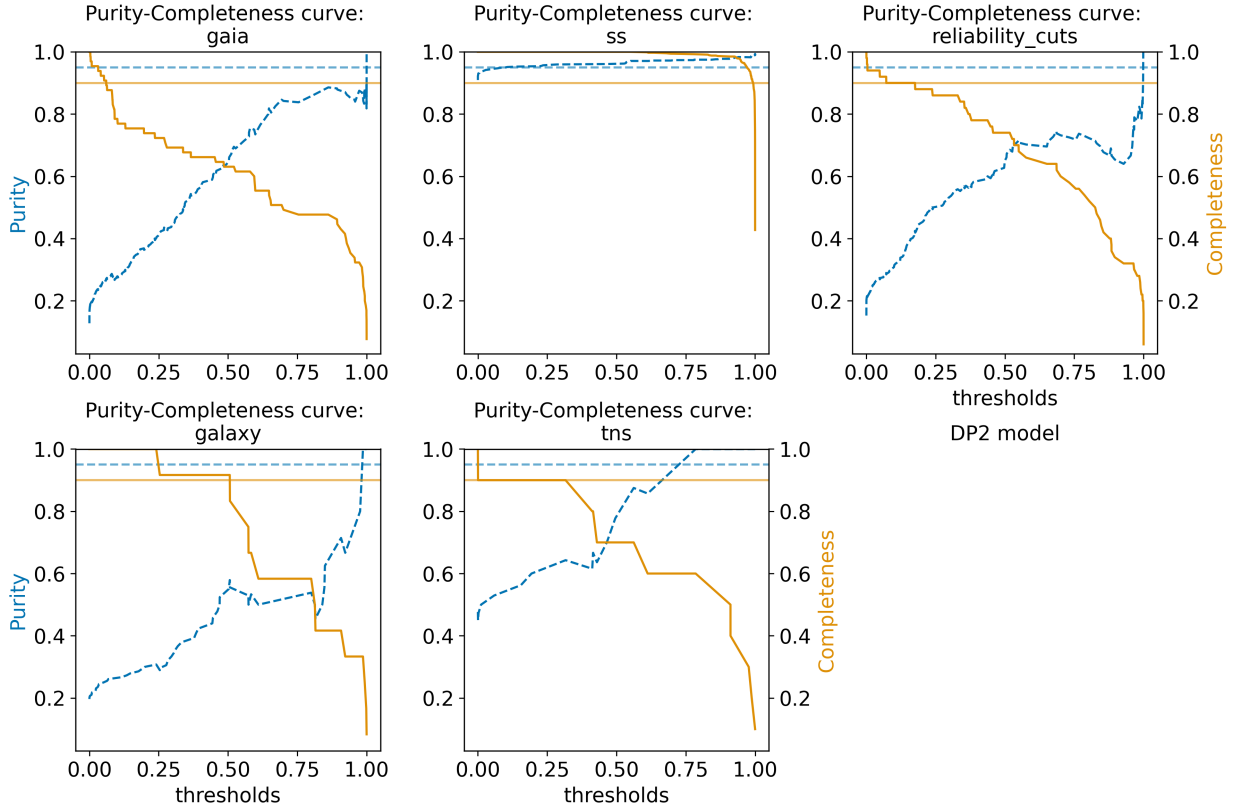


FIGURE 16: Purity and completeness per data origin for test data without fake injection (Zooniverse labels).

Observatory funding comes from private donations, grants to universities, and in-kind support from LSST-DA Institutional Members.

B References

[RTN-111], AlSayyad, Y., O'Mullane, W., 2026, *Data Preview 2: Definition and planning*, Technical Note RTN-111, NSF-DOE Vera C. Rubin Observatory, URL https://rtn-111.lsst.io/

[LSE-30], Claver, C.F., The LSST Systems Engineering Integrated Project Team, 2018, *Observatory System Specifications (OSS)*, Systems Engineering Controlled Document LSE-30, NSF-DOE Vera C. Rubin Observatory, URL https://ls.st/LSE-30

[LSE-61], Dubois-Felsmann, G., Jenness, T., 2019, *Data Management System (DMS) Requirements*, Systems Engineering Controlled Document LSE-61, NSF-DOE Vera C. Rubin Observatory, URL https://lse-61.lsst.io/, doi:10.71929/rubin/2587200

[DMTN-102], Graham, M.L., Bellm, E.C., Guy, L.P., et al., 2024, *LSST Alerts: Key Numbers*, Data Management Technical Note DMTN-102, NSF-DOE Vera C. Rubin Observatory, URL https://dmtn-102.lsst.io/, doi:10.71929/rubin/2997858

Inada, A., Sako, M., Acero-Cuellar, T., Bianco, F., 2026, AJ, 171, 205 (arXiv:2508.16844), doi:10.3847/1538-3881/ae38d8, ADS Link

LSST Dark Energy Science Collaboration, Abolfathi, B., Armstrong, R., et al., 2021, arXiv e-prints, arXiv:2101.04855 (arXiv:2101.04855), doi:10.48550/arXiv.2101.04855, ADS Link

Marshall, P.J., Verma, A., More, A., et al., 2016, Monthly Notices of the Royal Astronomical Society, 455, 1171

Paszke, A., Gross, S., Massa, F., et al., 2019, Advances in neural information processing systems, 32

Rimoldini, L., Holl, B., Gavras, P., et al., 2023, Astronomy & Astrophysics, 674, A14

Transient Name Server, IAU Supernova Working Group, URL https://www.wis-tns.org

[RTN-095], Vera C. Rubin Observatory Team, 2026, *The Vera C. Rubin Observatory Data Preview 1*, Technical Note RTN-095, NSF-DOE Vera C. Rubin Observatory, URL https://rtn-095.lsst.io/, doi:10.71929/rubin/2570536

C Acronyms

Acronym	Description
AP	Alert Production
APDB	Alert Production DataBase
AST	NSF Division of Astronomical Sciences
AURA	Association of Universities for Research in Astronomy
Adam	Adaptive Moment Estimation
B	Byte (8 bit)
CNN	Convolutional Neural Network

COSMOS	Cosmic Evolution Survey
DC2	Data Challenge 2 (DESC)
DE-AC02	Department of Energy contract number prefix
DIA	Difference Image Analysis
DMSR	DM System Requirements; LSE-61
DMTN	DM Technical Note
DP1	Data Preview 1
DP2	Data Preview 2
DRP	Data Release Processing
LSE	LSST Systems Engineering (Document Handle)
LSST	Legacy Survey of Space and Time
LSST-DA	LSST Discovery Alliance
LSSTCam	LSST Science Camera
LSSTComCam	Rubin Commissioning Camera
ML	Machine Learning
OSS	Observatory System Specifications; LSE-30
PSF	Point Spread Function
RTN	Rubin Technical Note
SLAC	SLAC National Accelerator Laboratory
SNR	Signal to Noise Ratio
SS	Subsystem Scientist
TNS	Transient Name Server