Vera C. Rubin Observatory
Data Management

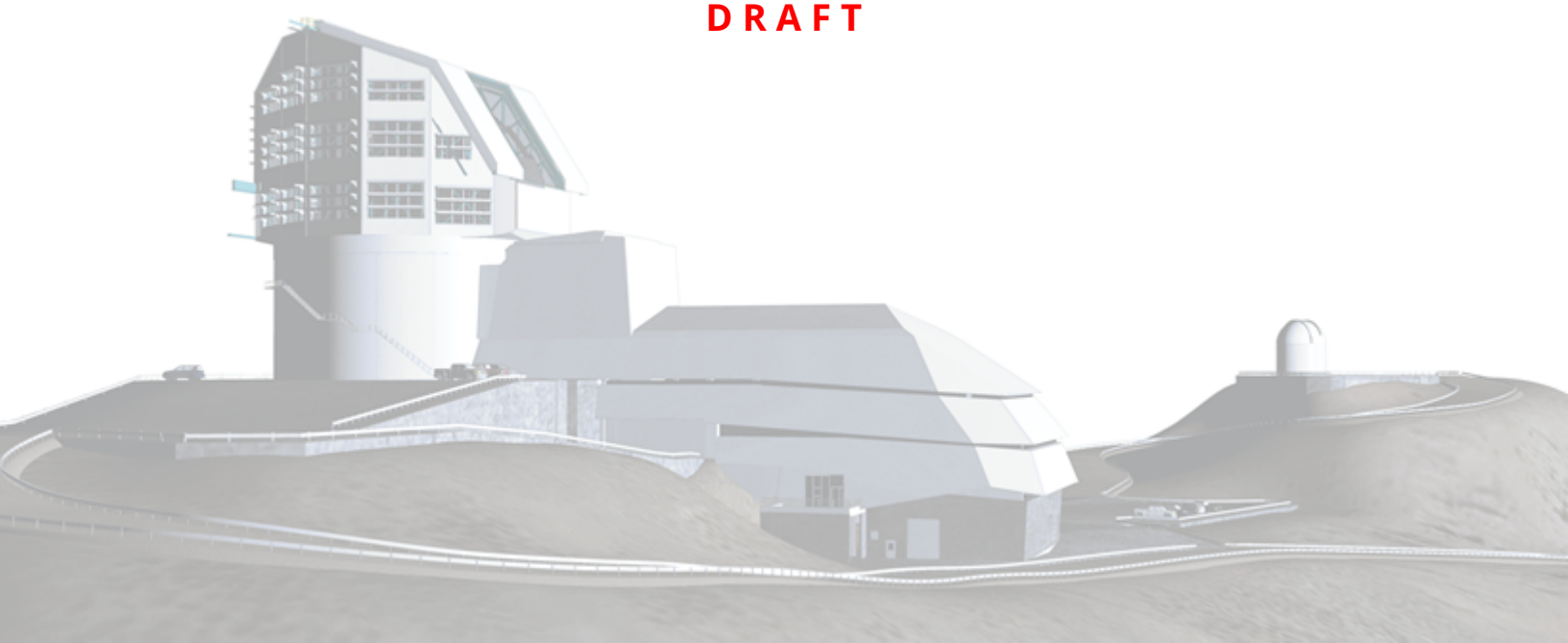
Performance of Machine-Learned Reliability Scoring for Image Differencing

Tatiana Acero-Cuellar, Federica Bettina Bianco, Eric C. Bellm

DMTN-337

Latest Revision: 2026-04-17

DRAFT



Abstract

Status of the reliability classifier (”real/bogus”)

Change Record

Version	Date	Description	Owner name
1	YYYY-MM-DD	Unreleased.	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 in the document *LSE-30 define the Difference Source Spuriousness Threshold - Transients | ID: OSS-REQ-0353* states 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, DMTN-102 "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, outputs of the Difference Image Analysis Pipeline, and returns the reliability score (Real/Bogus score), a number between 0 and 1. This is a supervised machine learning model, where the labels for each triplet of postage stamps are either Real or Bogus. The labels are gathered with a combination of fake injection, labels obtained by analyzing the classification given by volunteers, and also by Rubin Observatory members at a Zooniverse citizen science project called Rubin Difference Detectives. This note reports the performance of that classifier.

The set of weights that generate the reliability score has been updated three times; for each update, the model architecture remains the same, but a different dataset was used. Each version of the reliability model is detailed in Table 1.

Model version number	Butler collection name	AP deployment date	AP reliability cutoff	DRP usage
0.1	tac_cnn_comcam_2025-02-18		0.1	DP1 model
0.2	tac_cnn_lsstcam_2026-02-13		0.1 (before Feb 24, 2026) 0.5 (after Feb 24, 2026)	
0.3	tac_cnn_lsstcam_2026-02-26		0.5	DP2 model

TABLE 1: Model version

2 Description of the Machine Learning model for reliability score

The following section describes the Convolutional Neural Network architecture used to train the postage stamps and predict reliability scores. We developed a relatively simple model:

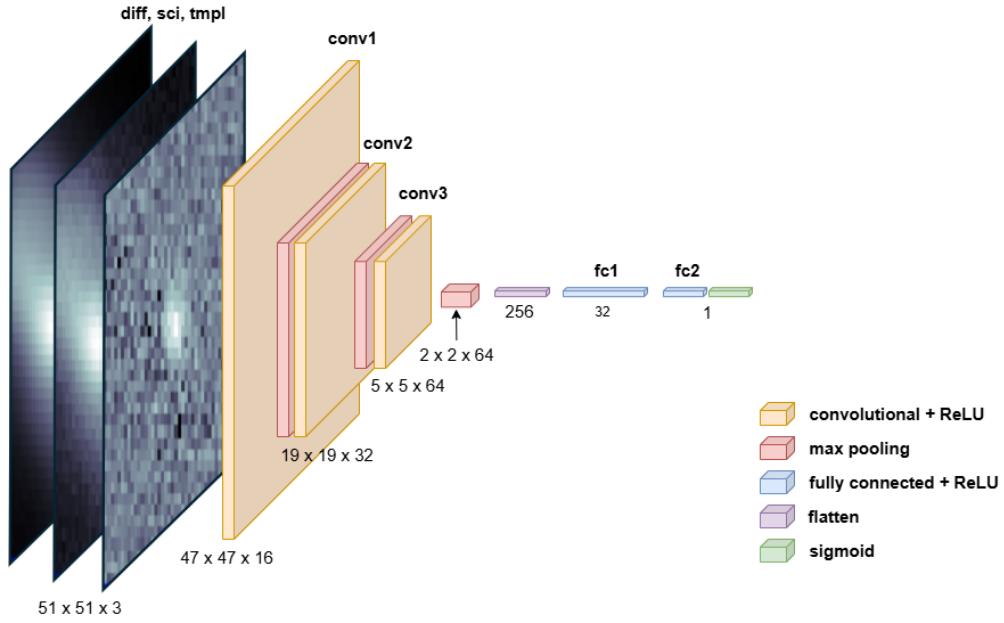


FIGURE 1: CNN architecture of the Machine Learning model that generates a reliability score for every DIASource. The input data is the template, science, and the difference image, each of size 51×51 pixels, and they are centered on the detected sources. The input array has a shape of (3, 51, 51). The 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.

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 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 was 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.

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 each model update.

3.1 Data for DP1 model: LSSTComCam data

The model was initially trained with 89,066 Real and the same amount of Bogus, extracted from DC2 simulations, plus random injections of stars to increase the number of Real. Once the LSSTComCam data were available, the model was fine-tuned on a subset of the data containing 183,046 sources with PSF injections. The fine-tuned model was the one used to generate reliability scores for (Vera C. Rubin Observatory Team, RTN-095, DP1).

3.2 Data for v0.2 and DP2 model

DP2 (v0.2) training partially (totally) relies on real (without fake injection) data and Real/Bogus labels obtained by the Zooniverse community through the 'Rubin Difference Detectives' project. The fine-tuned model was the one used to generate reliability scores for DP2.

Name	Num sources	Pipeline	Type
tns_ap	219	AP	TNS match
rel_cuts_drp33_gt_05_lt_09	49,933	DRP 2025_33	reliability cut
tns_drp33	855	DRP 2025_33	TNS match
tns_drp37	2984	DRP 2025_37	TNS match
gaia_drp37	89,410	DRP 2025_37	GAIA match
ss_drp37	88,058	DRP 2025_37	Solar System match
galaxy_ddf	10,932	AP	Galaxy match (extendedness=1)

TABLE 2: 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 an extendedness=1 given by Source catalogs.

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 community to label. The distribution of the DIA sources per catalog is described in Table 2.

The subject given by the community was an image composed of three postage stamps of size 51×51: template, science, and difference images, the DIA pixel output for each source. Volunteers were asked to answer 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.

For a subject to be retired, meaning volunteers can no longer see it, it must first have been classified as either Real or Bogus by two different volunteers, and they must have been in complete agreement. If this initial condition was not met, the subject was instead 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 Data analysis of 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 the same as that defined and implemented by Marshall et al. (2016). The labels obtained through this method are then the training/validation and testing labels to fine-tune the ML-reliability

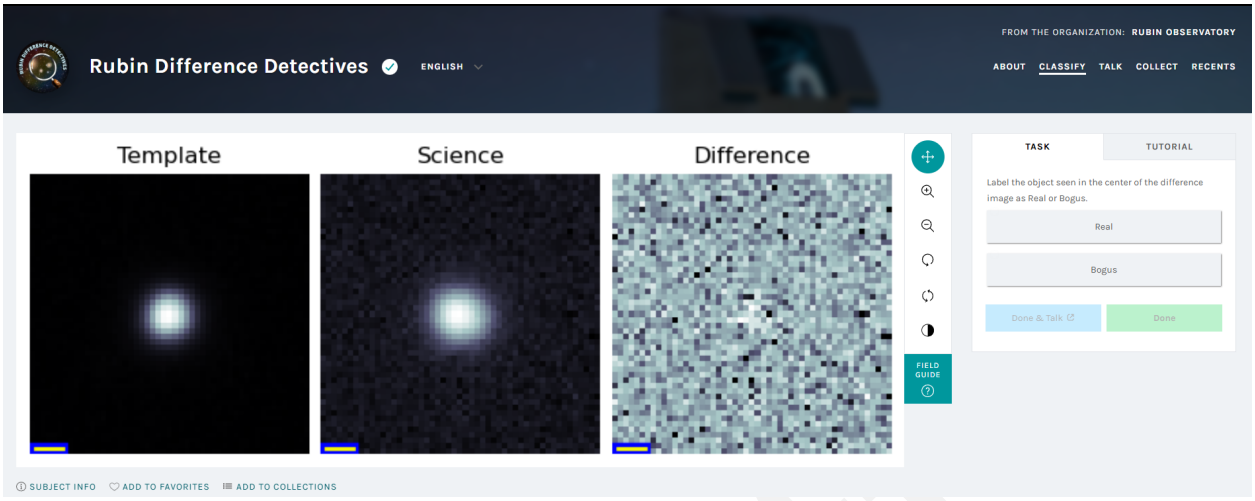


FIGURE 2: Zooniverse interface of the Rubin Difference Detectives project. The volunteers were shown a subject, which is an image with the template, science, and the difference postage stamps, and they were asked to select whether the source in the difference image is Real or Bogus.

model. In their work, they implemented a probabilistic classification for each subject (Pr(Real): probability of the DIA source to be a Real astrophysical object); the classifier was based on all the classifications given to the subject. Their probabilistic methodology depended on understanding how the volunteers responded to the reference data set (subjects where the ground truth was known). 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) or the DIA sources classified by the Rubin-ML-Reliability team, 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-ML-Reliability team. We note here that we only consider volunteers who provided labels for the reference set. From 3750 volunteers, only 2476 encountered a DIA source of the Rubin-ML-Reliability team reference data set (2792 when considering both reference datasets).

Since the Pr(Real) per DIA source returned a float between 0 and 1, to determined 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 True Positives (TP) and True Negatives (TN) ratios obtained by comparing volunteer classification and reference data labels. In general, the more labels a user makes, the higher the TP and TN ratio. Ideally, the TP and TN ratio should be similar, so that the user can distinguish both Real and Bogus sources.

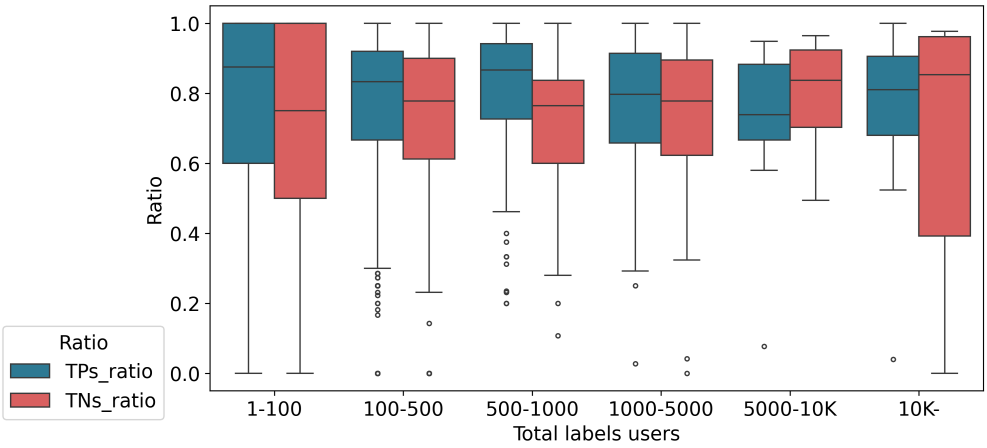


FIGURE 3: True Positives (TP) and True Negatives (TN) ratios obtained by comparing volunteer classification and reference classification, per total classifications made by the user. In general, the more classifications, the better the TP and TN ratio. There are some outliers where the volunteers are very optimistic and classify everything as Real, or very pessimistic and classify everything as Bogus.

The distribution of Real/Bogus for the reference data changes depending on the type of source, 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 and Figure 4 showed that low SNR tends to be classified as Bogus, regardless of the type of source (gaia, galaxy, reliability_cuts, ss, or tns).

After applying the Bayesian approach (Marshall et al. (2016)) to the Zooniverse classification, the data distribution is described in Table 3.

The total number of Real and Bogus after the classification analysis is unbalanced; for the purpose of training a binary classification machine learning model, the data sets need to be balanced between classes. The analysis and results shared here are based on a balanced

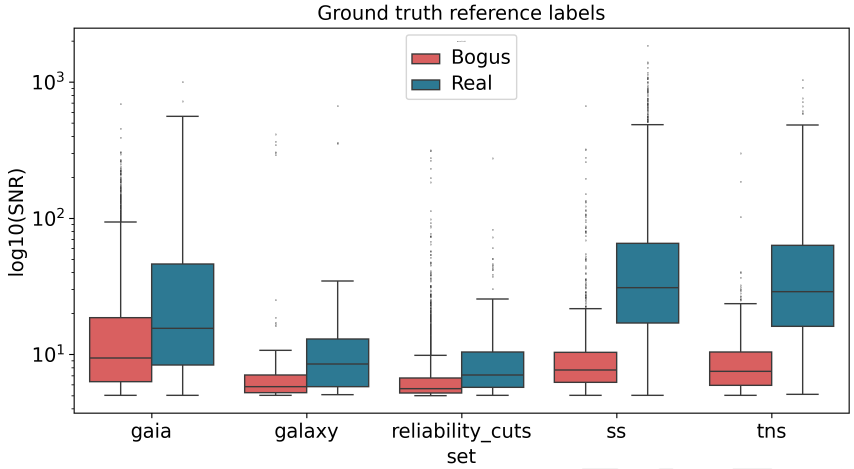


FIGURE 4: Signal to Noise Ratio distribution for each catalog. for DIA sources labeled by experts

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 3: Final count of Real and Bogus obtained after determining the probability of the source being Real given the classifications provided by the Zooniverse volunteers.

data set. The distribution of the Real and Bogus by data origin, as explained in Table 3, after analyzing the Zooniverse classifications, is shown in Figure 5.

The construction of the ground truth labels for fine-tuning the model for v0.2 and DP2 models relied on the Zooniverse analysis obtained by only considering the reference data set of the Rubin-ML-Reliability team (see Table 3). Additionally, only detections with high-confidence labels (stable) were used. In total, 13,178 sources were used to fine-tune the DP1 model: 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 DP2, a fraction of the dataset used for DP1 model (LSSTComCam fakes) was also used. For v0.2 the training relies 100% on the Zooniverse data.

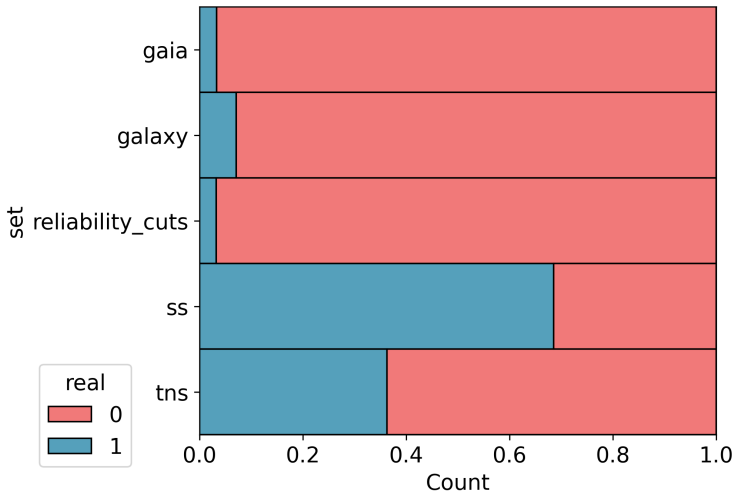


FIGURE 5: Distribution of Real and Bogus per origin of the data for data used for training/-validation/testing of the ML-reliability model. The data was grouped by the catalog used for cross-matching: tns_ap, tns_drp33, and tns_drp37 are all presented under 'tns', and the same for data obtained after cross-matching with the GAIA catalog. Distribution of each data origin for each class, Real (1) and Bogus (0). Almost 80% of the Real objects correspond to Solar System objects ('ss'), 10% correspond to variable objects (GAIA catalog), and the other 10% is distributed between transients (TNS catalog), objects with extendendness = 1 ('galaxy'), and Real found by the DP1 model ('reliability_cuts'). On the other hand, almost 60% of the Bogus objects correspond to matches with the GAIA catalog, mainly 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.

3.3 Data for Future models

The construction of the ground truth labels for fine-tuning the model relied on the Zooniverse analysis obtained by considering the training data set of both the Oxford team and the Rubin-ML-Reliability team (see Table 3). Additionally, we include fakes from more recent injection runs.

4 Performance

The distribution of the reliability score for a random set of DIA sources is shown in Figure 6. Besides the peaks at 0 and 1, each model presents a third peak at ~ 0.5 for DP1 model, and ~ 0.3 for both the DP2 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 at the moment, and to mitigate the impact on APDB, for each model, a reliability value cutoff is implemented. The cutoff is determined by analyzing the behavior of the model against properties of the DIA sources, and behavior on fakes, e.g. Figure 7, Figure 8, Figure 9. The idea is to minimize the amount of Bogus that reaches the Brokers and the scientific community. The cutoff values are 0.1, 0.5, and 0.5 for DP1 model, v0.2, and DP2 model, respectively.

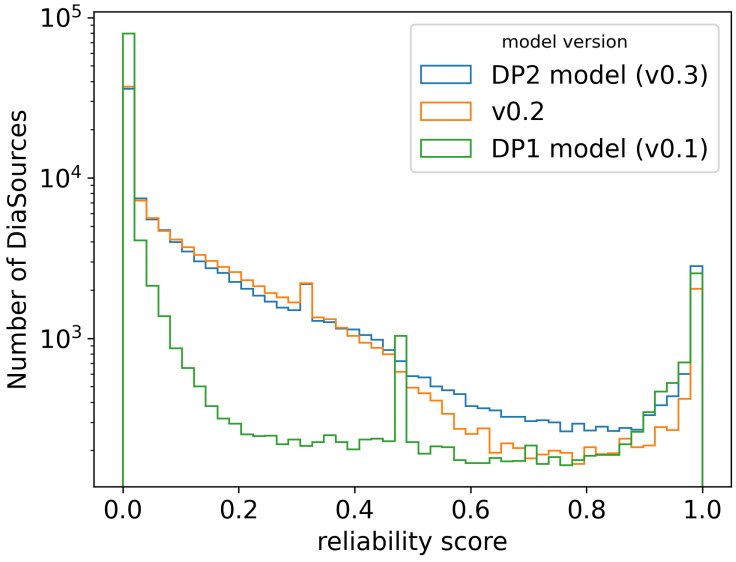


FIGURE 6: Distribution of the reliability score for each model version.

The behavior of each model against SNR for a fake injection run outside of the training data set is shown in Figure 7. DP1 model was trained only using fakes; v0.2 was only trained with real data without fake injection, and the DP2 model was trained with both fakes and real data.

The ratio of the quantity psfFlux (Flux for Point Source model) and the respective error, against the reliability score for each model, is shown in Figure 8. In general, DIA sources with large psfFlux error have a reliability score smaller than ~ 0.5. For the DP1 model, DIA sources with negative flux were not considered as part of the training set, then the model assigned to all DIA sources with negative flux values a reliability score below 0.5.

Following the basic idea behind the Difference Image Analysis (DIA), where the flux of the DIA source at the difference image is the subtraction of the science flux from the template flux. We show these two values (the difference flux and the subtraction) in Figure 9. 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 within the diagonal in Figure 9. Deviation

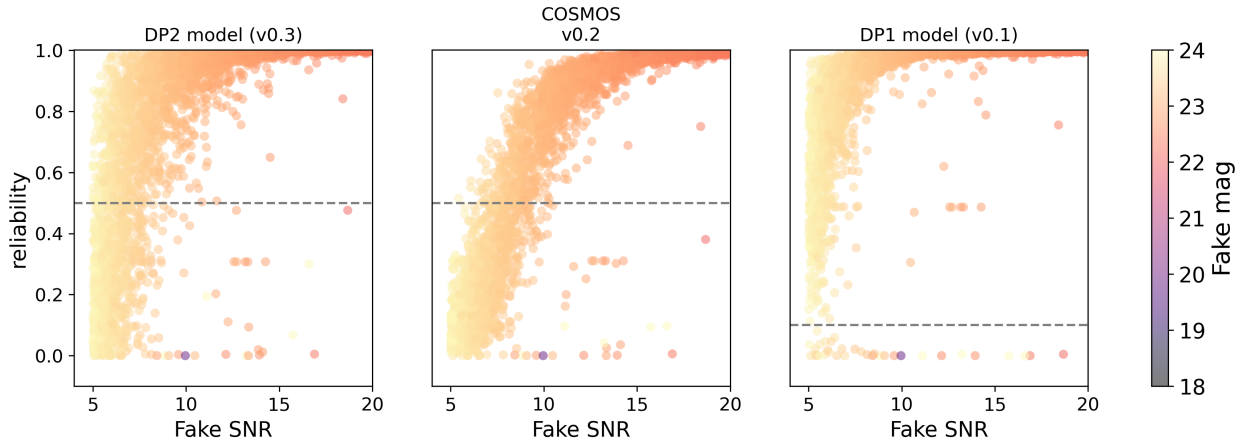


FIGURE 7: Reliability score against SNR for each model version for a set of fakes DIA sources from an injection run outside of the training set.

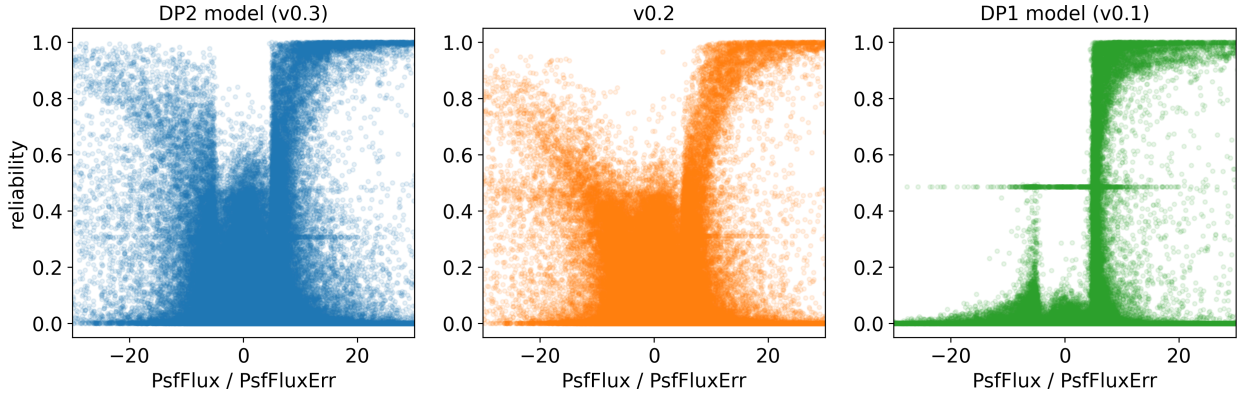


FIGURE 8: psfFlux/psfFluxErr vs reliability for each model version for a random set of DIA sources.

from that diagonal may indicate some deviation from this basic idea, and therefore a potential Bogus object (low reliability scores). Figure 10 shows the same plot but with the reliability cutoff value applied.

More details about the performance for each version are described in the following sections.

4.1 DP1 model (v0.1)

The DP1 model achieved a purity of 98.90% and completeness of 97.00% in the test dataset. On the LSSTComCam test set, the model achieved an accuracy of 98.06%, purity of 97.87%, and completeness of 98.27%. The weights up to this point were the ones used to generate

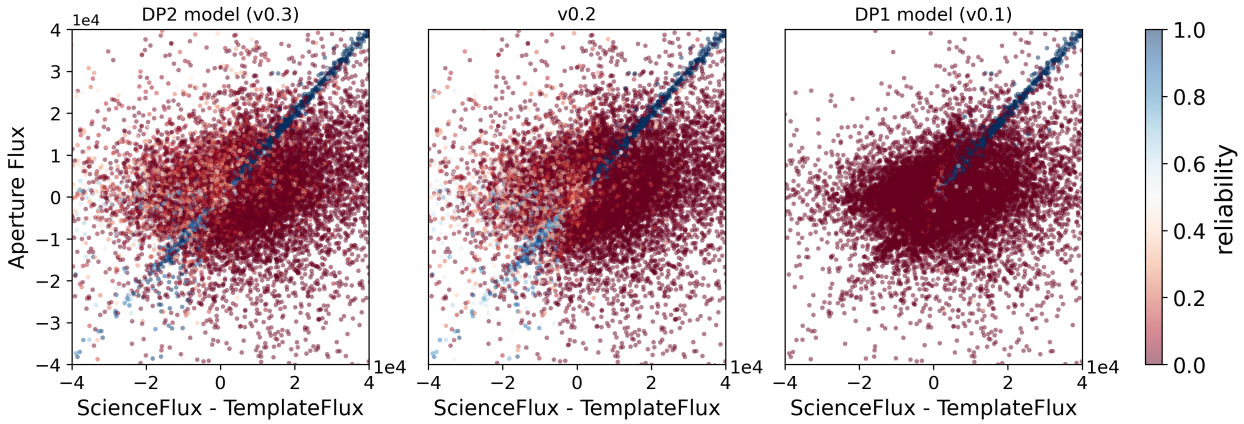


FIGURE 9: ScienceFlux - TemplateFlux vs Aperture Flux colored by the reliability score for each model version for a random set of DIA sources.

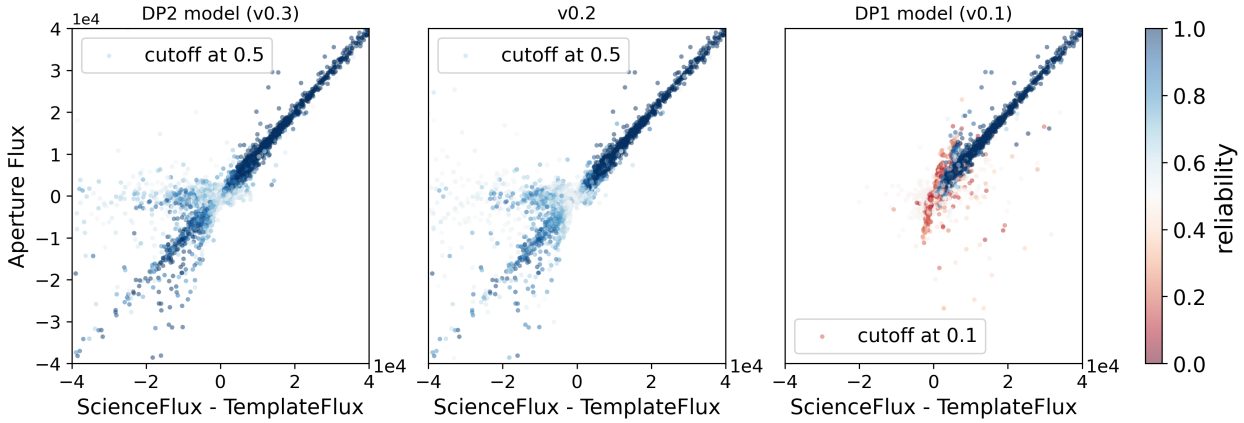


FIGURE 10: ScienceFlux - TemplateFlux vs Aperture Flux colored by the reliability score with a cutoff for each model version for a random set of DIA sources.

the reliability score on DP1 data (Vera C. Rubin Observatory Team, RTN-095, DP1). As reported in (Vera C. Rubin Observatory Team, RTN-095, DP1), and shown in Table 4, after applying a threshold of 0.5 to the reliability score, the purity of transient detections is high, but it had a limited impact on variable stars.

TABLE 4: 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

The DP1 model tested against a fake injection run outside of the training data set is shown in

Figure 11.

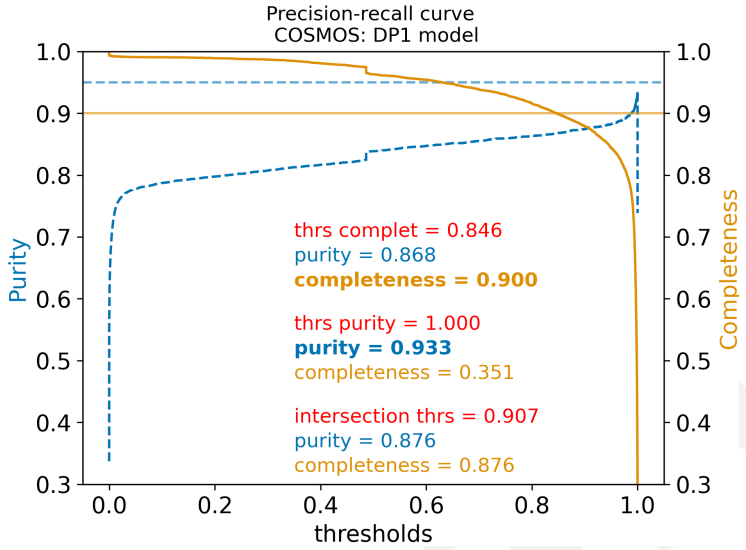


FIGURE 11: Performance of the DP1 model on fake sources injected into the COSMOS field.

4.2 model v0.2

The v0.2 model tested against a fake injection run outside of the training data set is shown in Figure 12.

4.3 DP2 model (v0.3)

The reliability score distribution per data catalog is shown in Figure 13.

The DP2 model reached a purity $\geq 95\%$ and completeness $\geq 90\%$ with threshold values between $0.808 \leq$threshold ≤ 0.833 for the test data without fake injections. For the LSSTComcam fakes test data, the model reached a purity $\geq 95\%$ and completeness $\geq 90\%$, which are obtained with values between $0.129 \leq$ threshold ≤ 0.899. Figure 14 shows the full purity and completeness curves.

Purity and Completeness per data origin for the real data (no fakes, only data with Zooniverse labels) test data is in Figure 15. The DP2 model tested against a fake injection run outside of the training data set is shown in Figure 16.

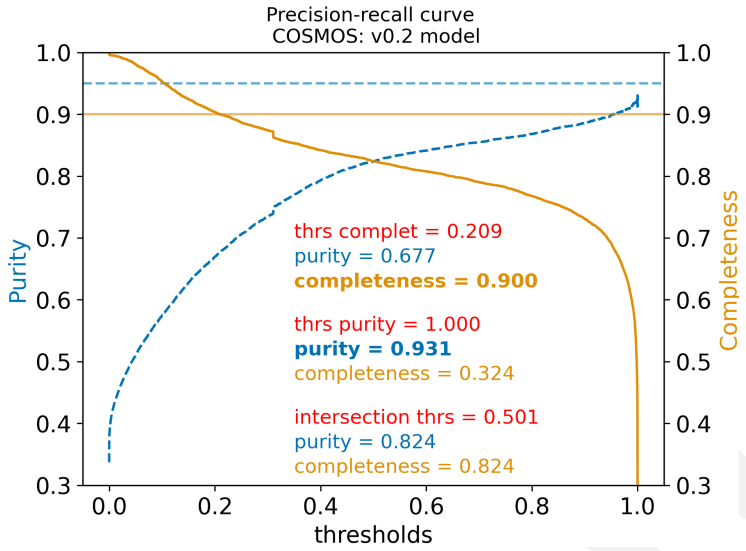


FIGURE 12: Performance of the v0.2 model on fake sources injected into the COSMOS field.

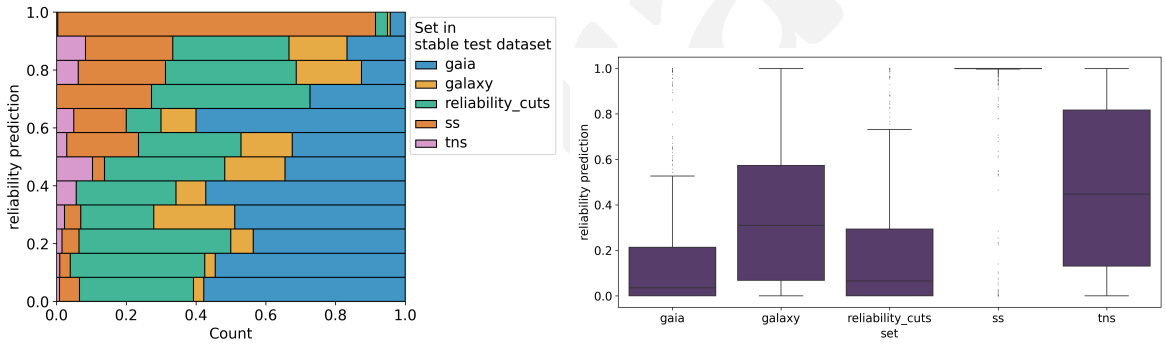


FIGURE 13: Reliability score on test data without fake injections.

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

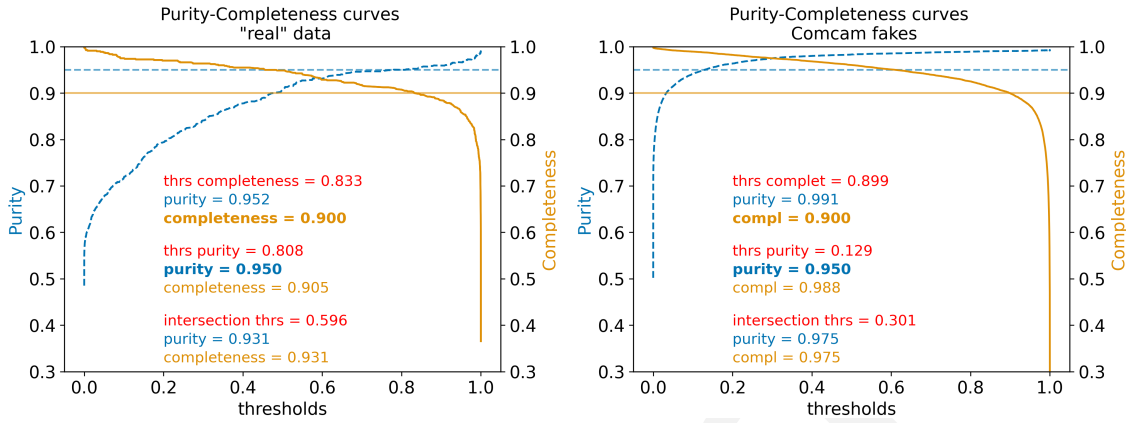


FIGURE 14: Purity and Completeness curves. *Left*: For the test data without fake injections (Zooniverse labels), at a threshold of 0.596, the purity (precision) and completeness (recall) are both 93.1%. *Right*: For the LSSTComcam fakes test data, at a threshold of 0.301, the purity (precision) and completeness (recall) are both 97.5%. Users can choose their own reliability threshold to trade off completeness vs. purity.

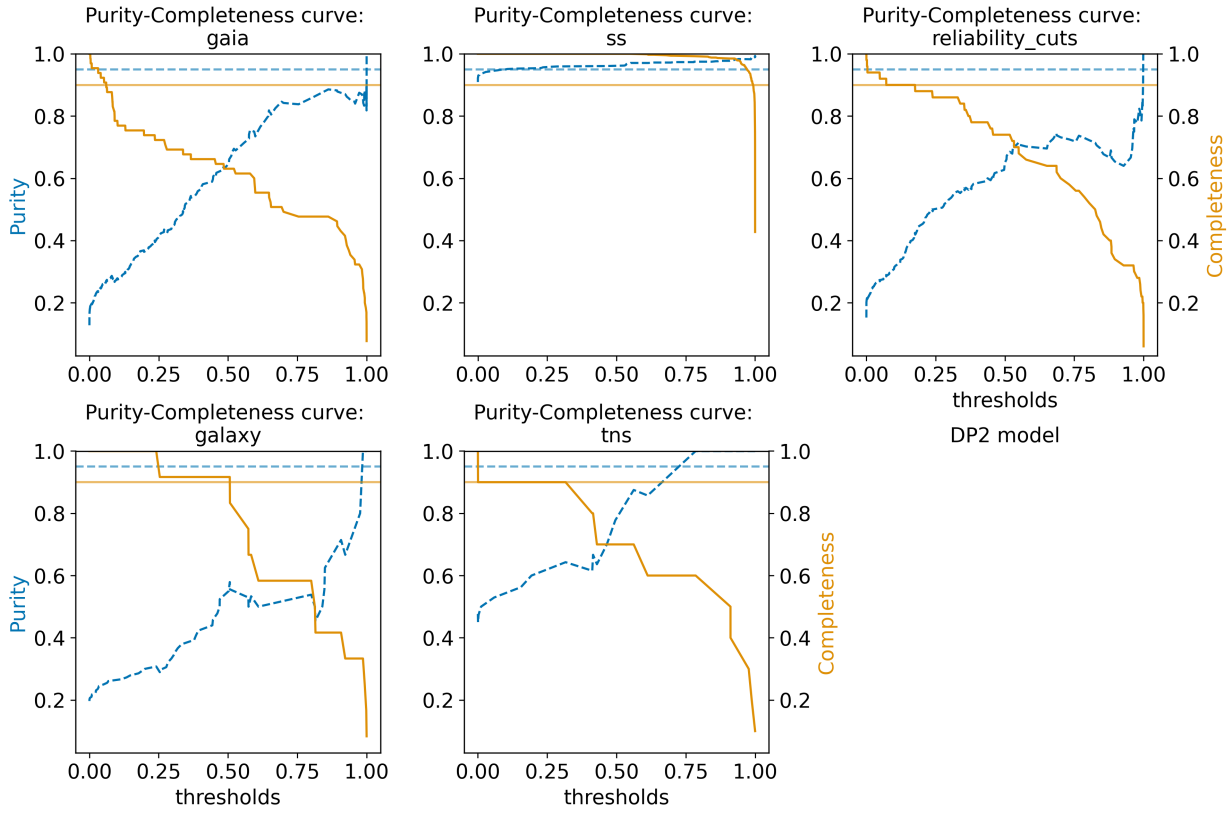


FIGURE 15: Purity and completeness per data origin for test data without fake injection (Zooniverse labels).

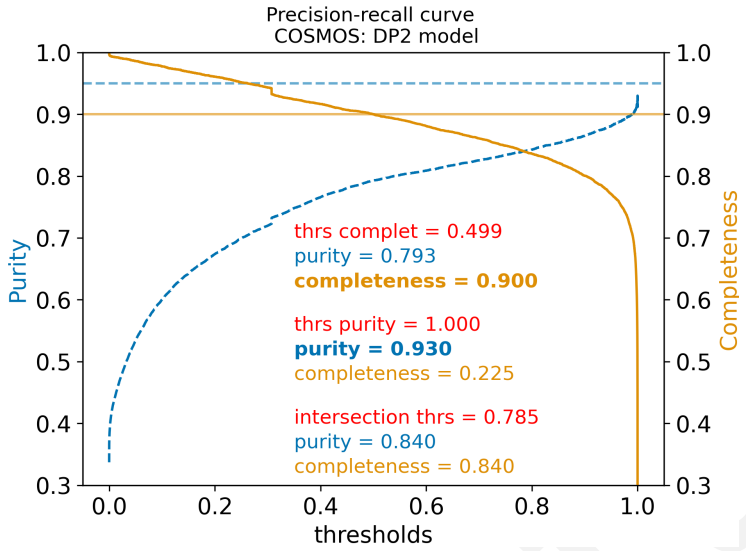


FIGURE 16: Performance of the DP2 model on fake sources injected into the COSMOS field.

B References

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

[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
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
DMTN	DM Technical Note
DP1	Data Preview 1
DP2	Data Preview 2
DRP	Data Release Processing
LSE	LSST Systems Engineering (Document Handle)
LSST-DA	LSST Discovery Alliance
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
TN	Technical Note
TNS	Transient Name Server