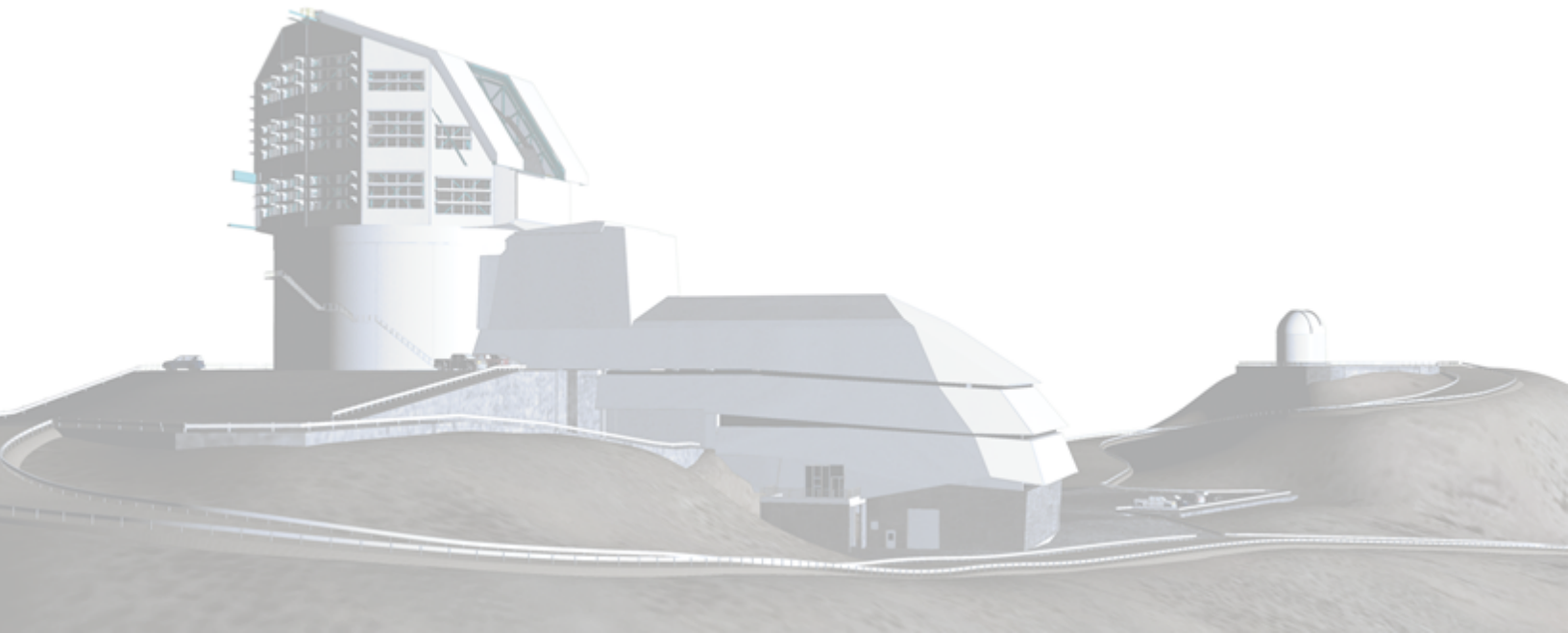

Vera C. Rubin Observatory
Data Management

Real-bogus classifier – status report

Nima Sedaghat

DMTN-272

Latest Revision: 2024-04-15

Abstract

We report the current status of the real-bogus classifier. The report starts with the design and technical aspects and continues to show quantitative and qualitative evaluations.

Change Record

Version	Date	Description	Owner name
1	YYYY-MM-DD	Unreleased.	Nima Sedaghat

Document source location: `https://github.com/lsst-dm/dmtn-272`

Contents

Real-bogus classifier – status report

1 Architecture

So far, we have run the experiments with three different architectures:

- An off-the-shelf architecture based on ResNet50 (He et al., 2016).
- An off-the-shelf architecture based on VGG6 (Simonyan & Zisserman, 2014).
- A custom architecture based on the encoder part of TransiNet, internally referred to as rbTN (Sedaghat & Mahabal, 2018).

The former two (ResNet50 and VGG6) have been more extensively tested.

2 Data

We used the version of the DC2 dataset available as a butler repository, as the main source of our data.

We have chosen the tract 3080 for training and 4024 for test/validation. Figures 1 , 2 show the distributions of exposure (≃visit) and detector numbers in the raw images currently captured (simulated) in each of the chosen tracts. The detector indices are, neglecting the narrow peaks, roughly uniformly distributed, meaning that we have an almost equal number of raw images (roughly 300) for each detector/CCD in each of the chosen tracts. This is particularly important for the training phase as we do not want the model to be biased towards specific detector characteristics.

More details about the exact data-set used for each experiment is included in the experiment's subsection.

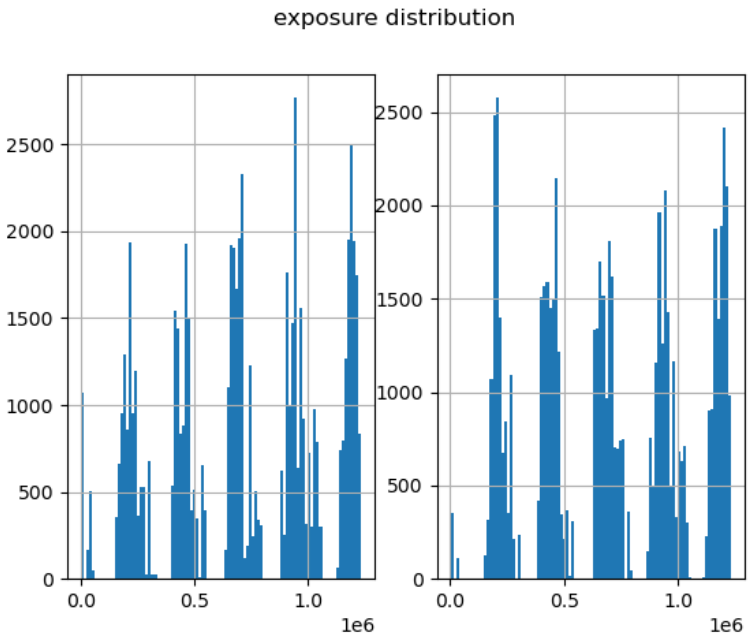

FIGURE 1: Distribution of exposure numbers of the existing raw images of each of the tracts. Left: tract #3080, right: tract #4024.

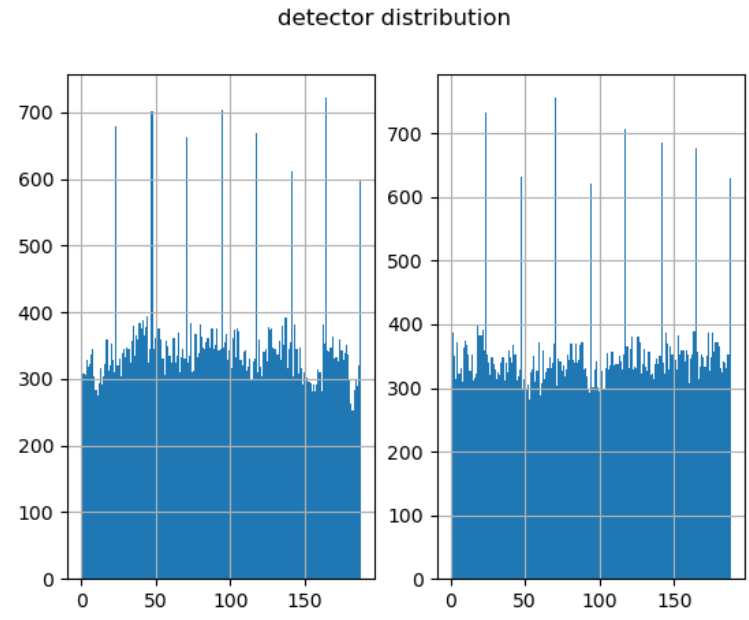

FIGURE 2: Distribution of the detector indices corresponding to the raw images of each of the tracts. Left: tract #3080, right: tract #4024.

Template images were created per-tract, in the sense that the data input to the APTemplate pipeline was constrained to a single tract at each run. Figure 14 illustrates an exemplar instance, where the effect of limiting the raws to a single tract appears as croppings in the output templates. This is not of a concern in this very application though, since the downstream tasks receive proper masks and will generate cutouts only of the regions from which enough informative data can be extracted.

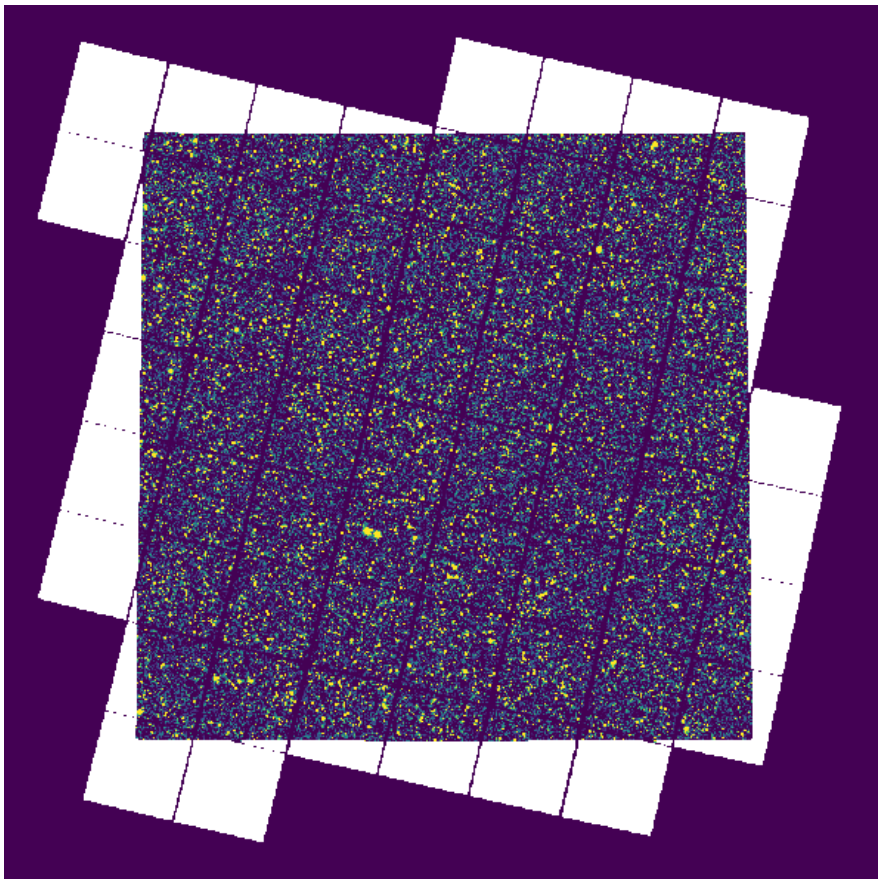

FIGURE 3: Templates generated based on raw images from a single tract. The white chessboard in the background is the spatial spread of the detectors (a single visit/exposure), whereas the area "full of stars" represents a single tract in the sky.

3 Evaluation Results

3.1 The "PCW 2022" network and data

The data-set used for training this model, DL Dataset v0.1.1, consists of 78222 triplets, out of which 2591 triplets (~ 3%) are `positives`.

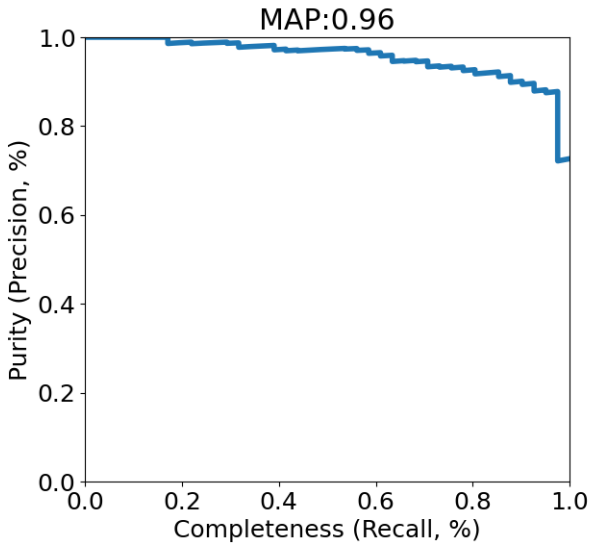

FIGURE 4: No sigma-thresholding applied on the input sources (diffim output)

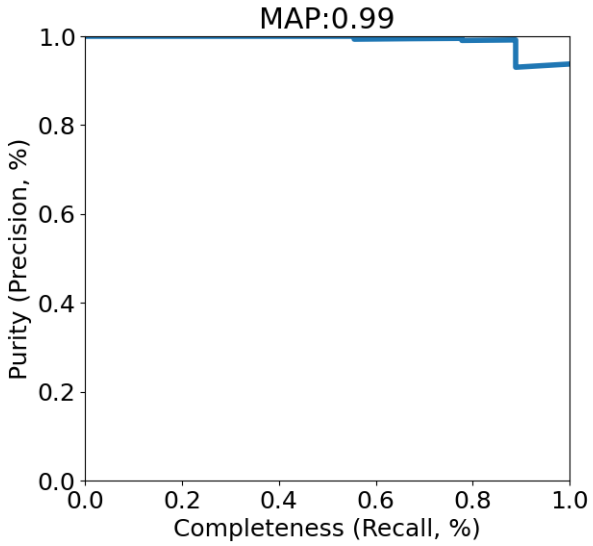

FIGURE 5: Only sources with $snr > 5\sigma$

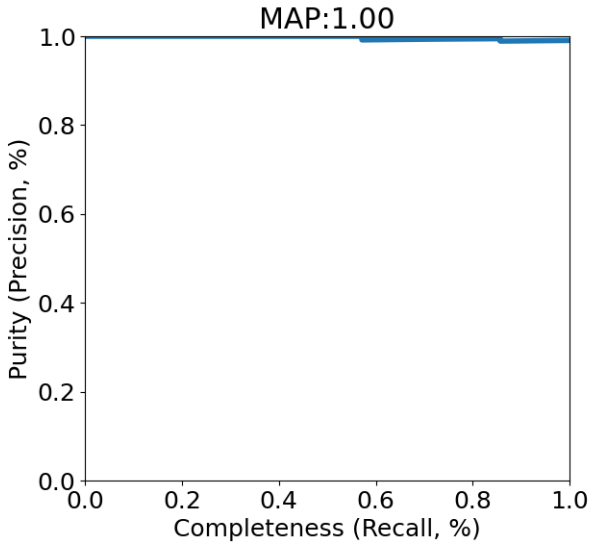

FIGURE 6: Only sources with $snr > 6\sigma$

3.2 New Dataset

The new dataset (DL v0.1.2) is essentially the same as the previous version (v0.1.1), only regenerated with a more recent version of the stack (`w_2023_38`). The number of images is relatively higher though, in the new version: there are `183750` triplets, out of which `10147` ($\sim 5\%$) are positives.

We used a $95\% - 5\%$ training-validation split, which leaves us with `9198` validation triplets (with `533` $\simeq 5\%$ positives). Below evaluation results are based on a snapshot of the network after 255K iterations (937 epochs).

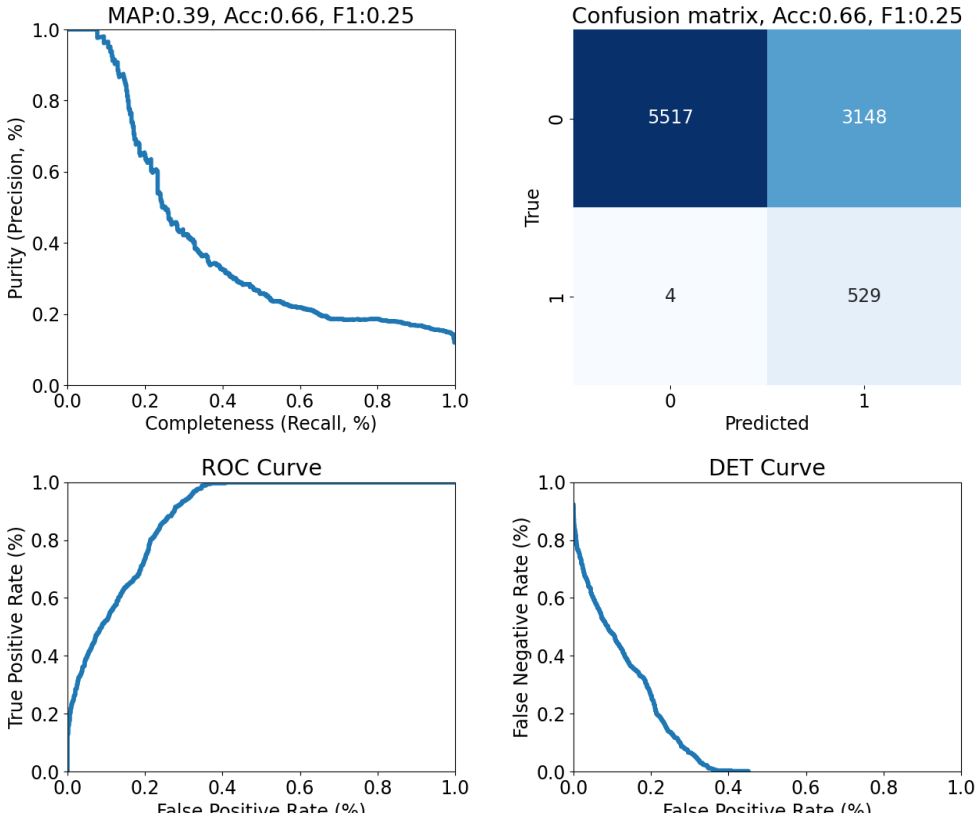

FIGURE 7: Evaluation results based on the whole dataset.

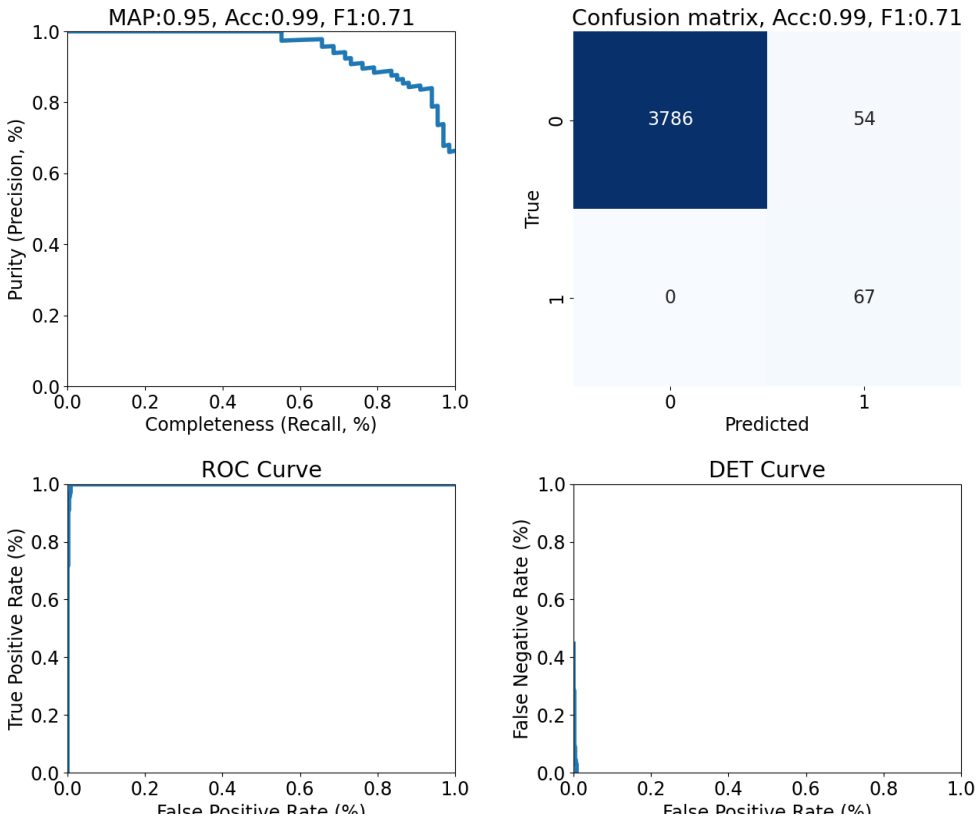

FIGURE 8: Only sources with $snr > 5\sigma$

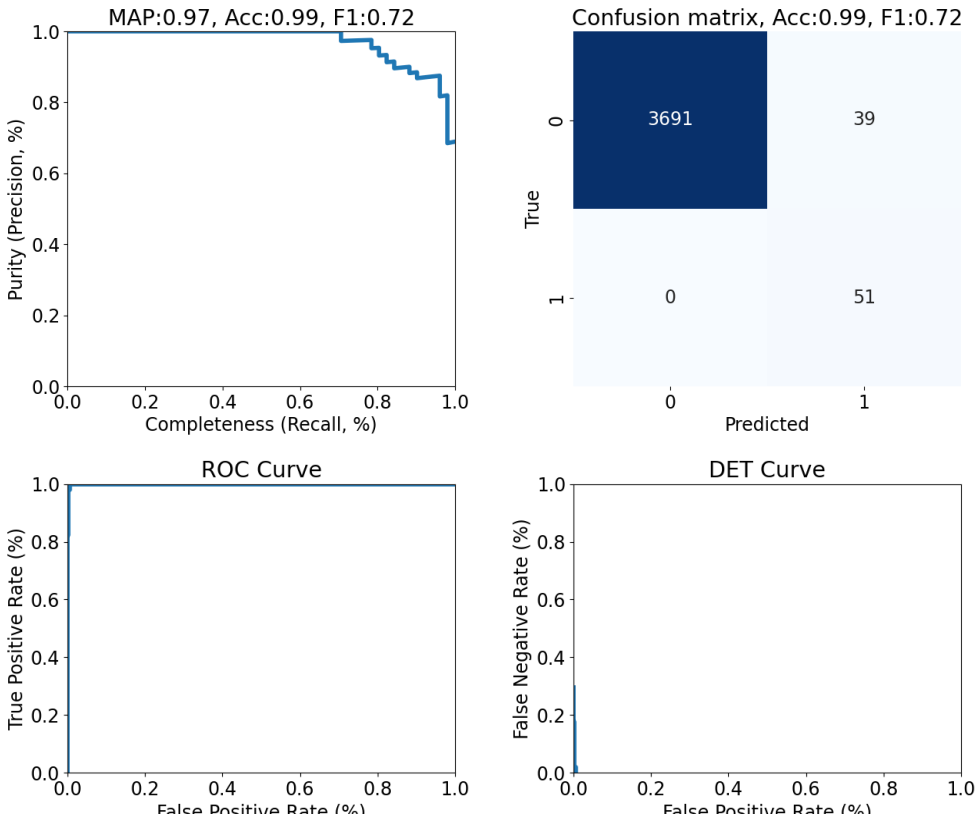

FIGURE 9: Only sources with $snr > 6\sigma$

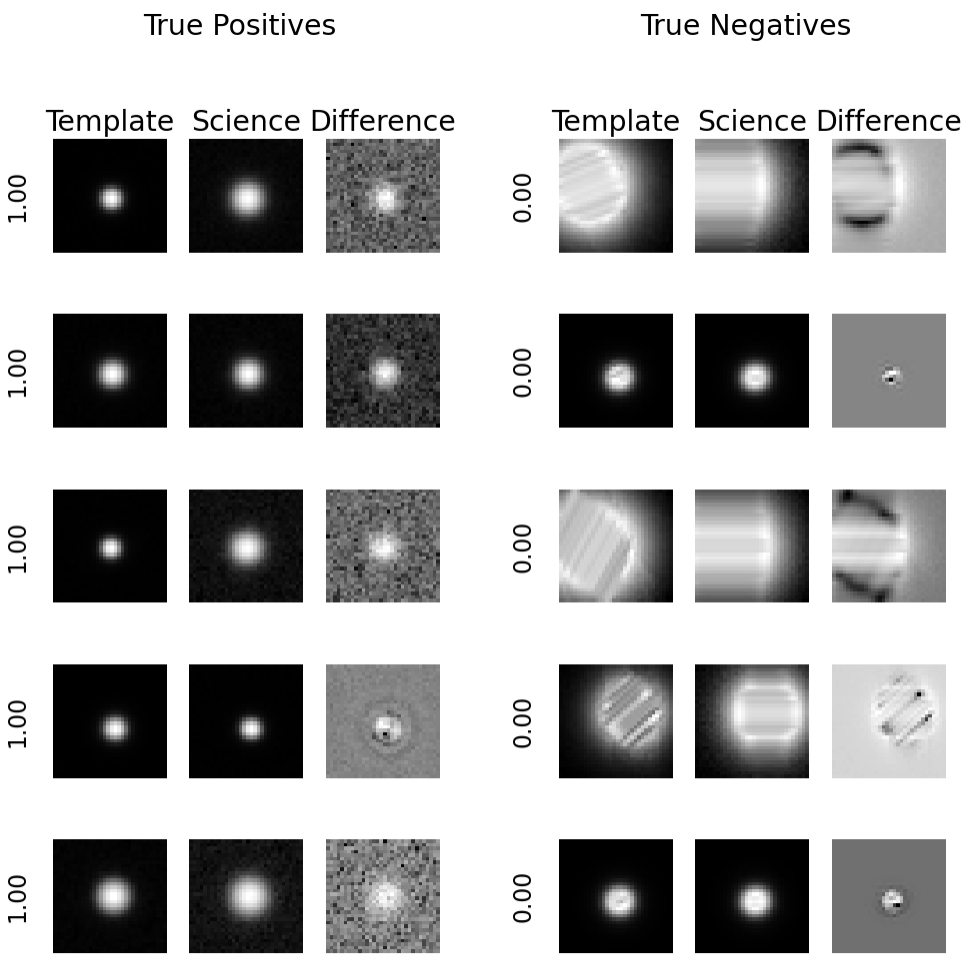

FIGURE 10: Exemplar top-scored triplets from the two “true” categories.

3.3 Training for Longer

Below you can see evaluation results of the same network as in Section 3.2, but after the network has been trained for 8M iterations (2934 epochs). Note the improvement in the results, but also refer to the notes in the next section (4) on how these numbers should be interpreted.

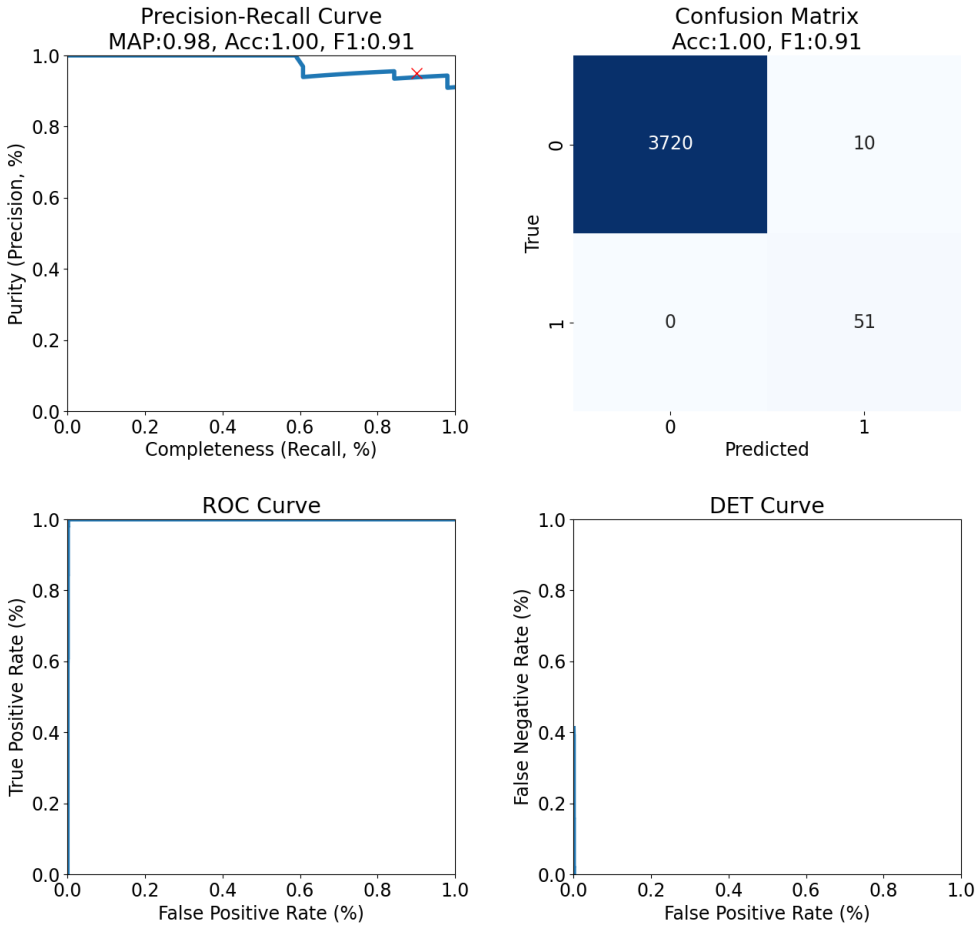

FIGURE 11: Only sources with $snr > 6\sigma$ (with weight-balancing). The red cross indicates the (0.9,0.95) point as defined in OSS-REQ-0353.

4 Summary

The quantitative results show that the simple preliminary binary classifier is close to meeting the requirements of the project – e.g. see Figure 11. Direct investigation of the only 10 incorrectly classified samples and taking consequent actions is of course possible, to push the performance of the classifier on precursor data even more – e.g. see DM-42867. Note, however, that this approach should be followed carefully and is not advised as a permanent solution: the more tailor-made the model gets for the precursor data at hand, the less comparable results we should expect from it when it is tested on real data – the already discussed issue of generalization.

A Evaluations with weight-balancing

Below plots show evaluation results @255000 iterations (comparable to Section 3.2) where a 20:1 weight has been applied on positive samples for generating the evaluation plots, to compensate for the class imbalance.

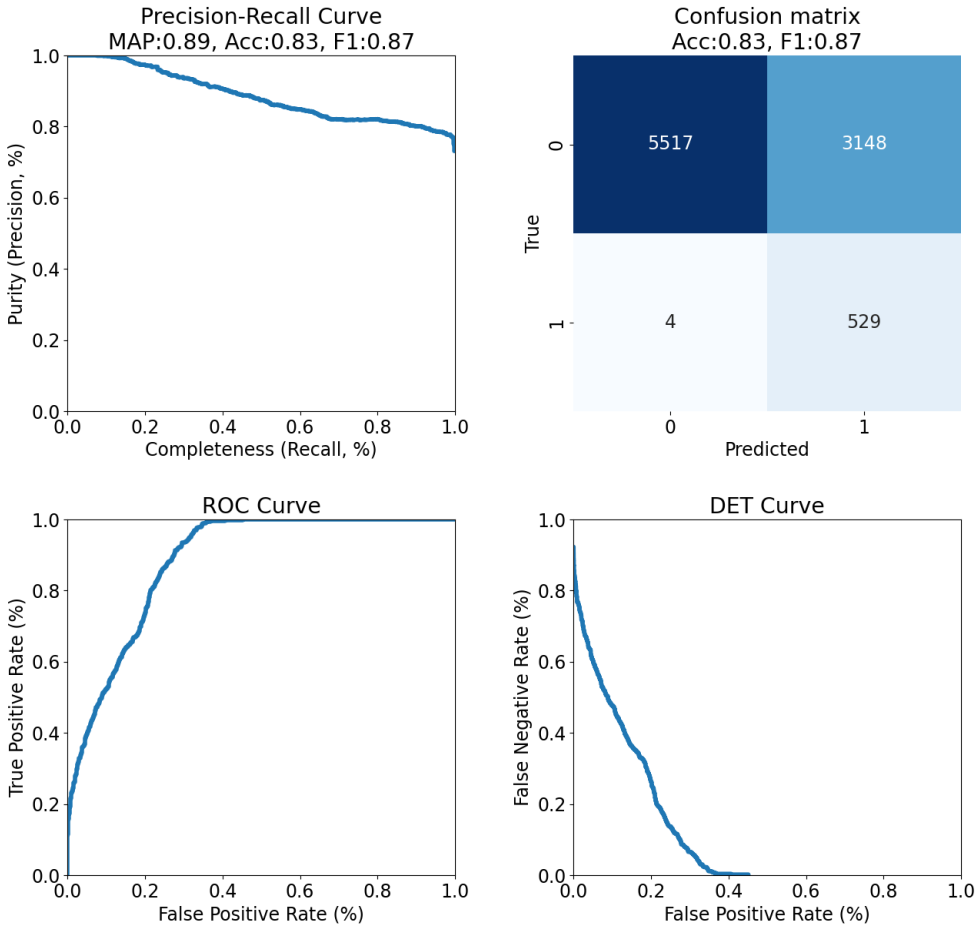

FIGURE 12: Evaluation results based on the whole dataset (with weight-balancing).

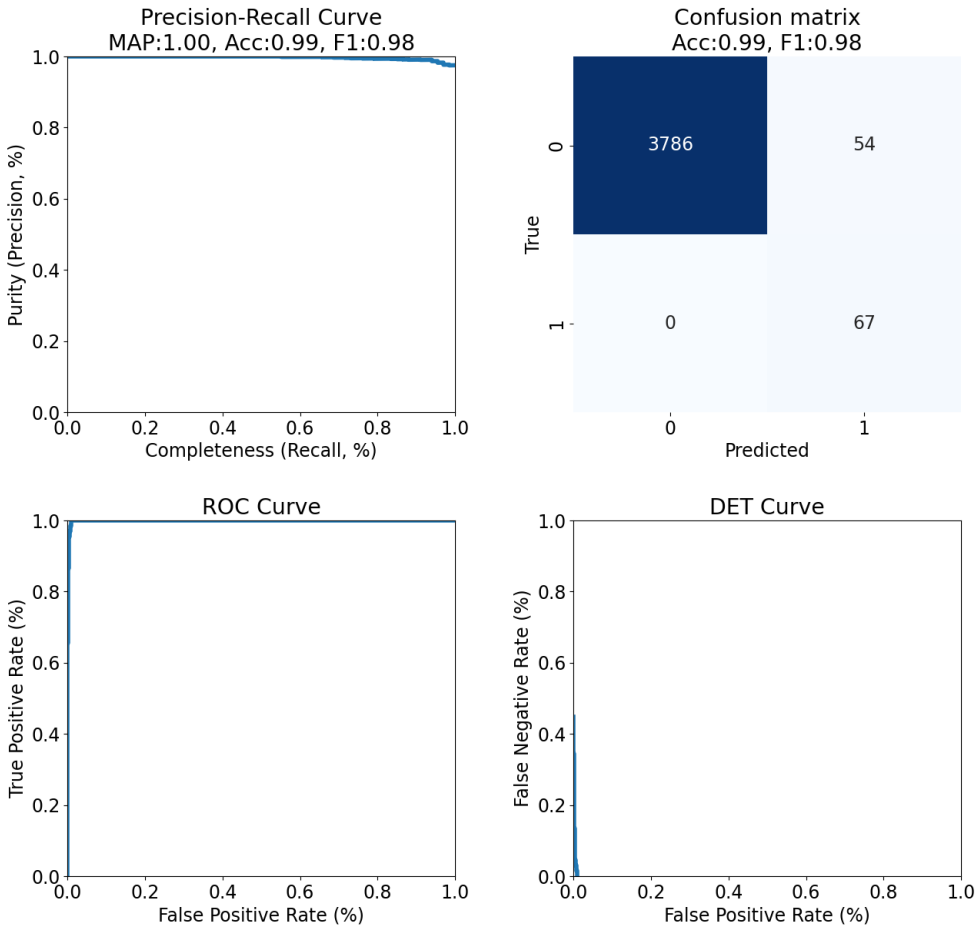

FIGURE 13: Only sources with $snr > 5\sigma$ (with weight-balancing)

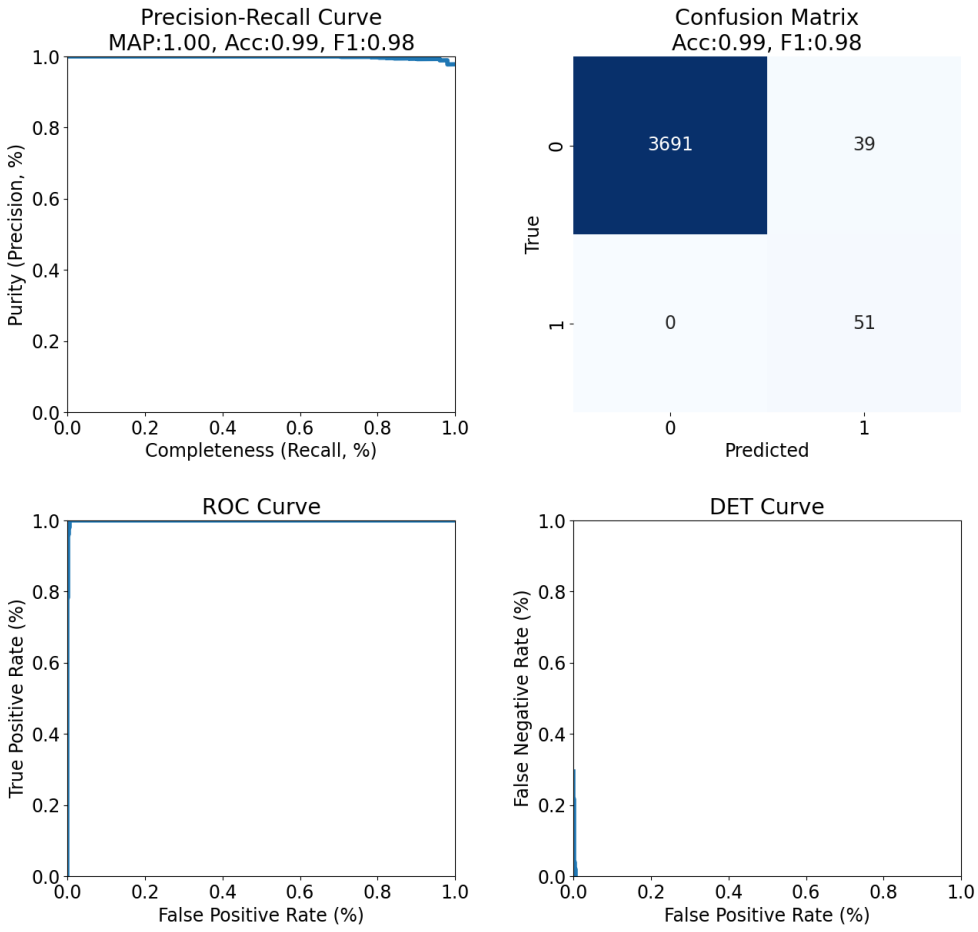

FIGURE 14: Only sources with $snr > 6\sigma$ (with weight-balancing)

B References

He, K., Zhang, X., Ren, S., Sun, J., 2016, In: Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778

Sedaghat, N., Mahabal, A., 2018, Publications of the Astronomical Society of the Pacific, 130, 114502

Simonyan, K., Zisserman, A., 2014, arXiv preprint arXiv:1409.1556

C Acronyms

Acronym	Description
CCD	Charge-Coupled Device
DC2	Data Challenge 2 (DESC)
DM	Data Management
DMTN	DM Technical Note
OSS	Observatory System Specifications; LSE-30
PCW	Project Community Workshop
TBA	To Be Announced