

Corresponding author(s): Julian A. Villalba

Last updated by author(s): 6/4/26

Machine Learning Checklist v 1.1

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form is intended to provide structure for consistency and transparency in reporting of works using or developing Machine Learning models. Some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

1. Availability and reproducibility of Code and Data

Please select all that apply regarding the availability of the data and code used in the study.

☐ Code will be included in a CodeOcean capsule.

☒ The source code is included in the submission or available in a public repository:

Custom scripts used for selected image quantification and cell detection are included with the submission.

Publicly available model files used for StarDist2D cell detection are available at: <https://github.com/qupath/models/tree/main/stardist>

☒ A compiled standalone version of the software is included in the submission or available in a public repository:

QuPath v0.6.0: <https://qupath.github.io/>

☐ A test dataset and instructions/scripts for replicating the results are included in the submission or available in a public repository:

Custom scripts are provided, but no test dataset is included because whole-slide image files are large and derive from a single fatal human case. WSIs can be provided to interested parties upon reasonable request. Source data are available from the corresponding author upon reasonable request, subject to institutional, privacy, and public health restrictions.

☐ A Readme file with instructions for installing and running the code is included in the submission or available in a public repository:

Brief instructions given at top of code files.

☒ The code is made available to reviewers during review.

☒ Pretrained models are used in the study and accessible through:

StarDist2D pretrained model for QuPath brightfield cell detection: <https://github.com/qupath/models/tree/main/stardist>

☒ Pretrained models are used in the study and are not accessible.

☒ The paper contains information on how to obtain code and data after publication.

2. Datasets

A. All data sources are listed in the paper.

☒ Yes

☐ No

B. The train, test and validation datasets are publicly available, and links/accession numbers have been provided in the manuscript or supplementary materials.

☐ Yes

☒ No

- C. We have reported and discussed potential dataset biases in the paper. Where applicable, appropriate mitigation strategies were used.
- ☒ Yes Yes, Potential dataset biases and mitigations relevant to the machine-learning image analyses are described
- ☐ No _____
- D. The data cleaning and preprocessing steps are clearly and fully described, either in text or as a code pipeline.
- ☒ Yes Yes. Data cleaning and preprocessing steps for the machine-learning image analyses are described in the
- ☐ No _____
- E. Instances of combining data from multiple sources are clearly identified, and potential issues mitigated.
- ☐ Yes _____
- ☒ No Not applicable. Multiple datasets were not combined for machine-learning model training. The QuPath pixel classifier was trained using project-specific manual annotations from representative ISH-stained tissue and then applied uniformly to annotated ROIs. Multimodal clinical, imaging, histopathological, molecular, and

3. Model and training

- A. What model architecture is the current model based on? Random trees (RTrees)
- B. A Model Card is provided¹.
- ☐ Yes
- ☒ No
- C. The model clearly splits data into different sets for training (model selection), validation (hyperparameter optimization), and testing (final evaluation).
- ☐ Yes
- ☒ No
- D. The method of data splitting (e.g. random, cluster- or time-based splitting, forward cross-validation) is clearly stated.
- ☐ Yes _____
- ☒ No No formal data splitting was performed. The classifier was trained using QuPath's built-in "Train pixel classifier" functionality with project-specific manual annotations. Classifier parameters are described in the Methods.
- E. The data splitting mimics anticipated real-world applications.
- ☐ Yes _____
- ☒ No Not applicable.
- F. The data splitting procedure has been chosen to avoid data leakage.
- ☐ Yes _____
- ☒ No Not applicable.

¹ <https://huggingface.co/docs/hub/model-cards>

G. The interpretability of the model has been studied and clearly validated.

- ☒ Yes Qupath classifiers are commonly used. Classifier outputs were analyzed and visually reviewed in the context
☐ No _____

4. Evaluation

A. The performance metrics used are described and justified in the paper.

- ☒ Yes Partially described in methods starting in line 903 (subsection POTV RNA density (POTVRNA-index) and
☐ No _____

B. Cross-validation of the results is included.

- ☐ Yes
☒ No

C. Community-accepted benchmark datasets/tasks are used for comparisons.

- ☐ Yes _____
☒ No None exist.

D. Baseline comparisons to simple/trivial models (for example, 1-nearest neighbour, random forest, most frequent class) are provided.

- ☐ Yes _____
☒ No Not applicable.

E. Benchmarks with current state-of-the-art are provided.

- ☐ Yes _____
☒ No Not applicable.

F. Ablation experiments are included.

- ☐ Yes _____
☒ No Not applicable.

G. The model has been tested on a fully independent dataset.

- ☐ Yes
☒ No

5. Computational resources

A. The paper contains information on hardware/computing resources that were used.

- ☐ Yes
☒ No

B. The paper includes information on the computational costs in terms of computation time, parallelization or carbon footprints estimates.

- ☐ Yes
☒ No