**Checklist for reporting ML-based science (beta)**

**About the checklist**

This checklist lists items that should be reported in a scientific study that uses machine learning (ML) methods. It is intended to accompany the paper or report that introduces an ML model: for instance, as an appendix or supplemental material. The checklist consists of 35 questions spread across 8 sections. For each item, either list the section name, section number, or page number in the paper where the item is reported, or justify why a given item is not filled out. Note that not all of these items need to be reported in the main text of the paper; they could be reported in an appendix or supplemental material.

Some items in the checklist could be hard to report for specific studies. For instance, including a reproduction script to computationally reproduce all results (2e.) might not be possible for studies performed on academic computing clusters or those which use private data that cannot be released. Instead of requiring strict adherence for each item, we suggest authors and referees decide which items are relevant for a study and where details can be reported better. The items in our reporting standards could be a helpful starting point.

Use the accompanying Guidelines for reporting ML-based science to see how each item can be filled out.

This is a beta version of our checklist. We are soliciting feedback and will continue to update the template. For feedback or questions, contact: sayashk@princeton.edu. The checklist starts on the next page. After filling it out, save it starting from that page.

**Authors**

Sayash Kapoor Inioluwa Deborah Raji

Thanh Hien Pham Michael Roberts

Christopher A. Bail Matthew J. Salganik

Emily Cantrell Marta Serra-Garcia

Odd Erik Gundersen Brandon M. Stewart

Jake M. Hofman Gilles Vandewiele

Jessica Hullman Arvind Narayanan

Michael Lones

Momin M. Malik

Priyanka Nanayakkara

Kenny Peng

Russell A. Poldrack

**Checklist for reporting ML-based science**

**Section 1: Study design**

1a. Population or distribution about which the scientific claim is made.

1b. Motivation for choosing this population or distribution (1a).

1c. Motivation for the use of ML methods in the study.

**Section 2: Computational reproducibility**

2a. Dataset used for training and evaluating the model along with link or DOI to uniquely identify the dataset.

2b. Code used to train and evaluate the model and produce the results reported in the paper along with link or DOI to uniquely identify the version of the code used.

2c. Description of the computing infrastructure used.

* Hardware infrastructure: CPU, GPU, RAM, disk space etc.
* Operating system.
* Software environment: Programming language and version, documentation of all packages used along with versions and dependencies (e.g., through a requirements.txt file).
* An estimate of the time taken to generate the results.

2d. README file which contains instructions for generating the results using the provided dataset and code.

2e. Reproduction script to produce all results reported in the paper[[1]](#footnote-0).

**Section 3: Data quality**

3a. Source(s) of data, separately for the training and evaluation datasets (if applicable), along with the time when the dataset(s) are collected, the source and process of ground-truth annotations, and other data documentation.

3b. Distribution or set from which the dataset is sampled (i.e., the sampling frame).

3c. Justification for why the dataset is useful for the modeling task at hand.

3d. The outcome variable of the model, along with descriptive statistics (split by class for a categorical outcome variable) and its definition.

*(The outcome variable is also known as the dependent variable, the target variable, the output variable or the predictor variable).*

3e. Number of samples in the dataset.

3f. Percentage of missing data, split by class for a categorical outcome variable.

3g. Justification for why the distribution or set from which the dataset is drawn (3b.) is representative of the one about which the scientific claim is being made (1a.). If it is not representative, note among threats to external validity (8a.)

**Section 4: Data preprocessing**

4a. Identification of whether any samples are excluded with a rationale for why they are excluded.

4b. How impossible or corrupt samples are dealt with.

4c. All transformations of the dataset from its raw form (3a.) to the form used in the model, for instance, treatment of missing data and normalization. Preferably through a flow chart.

**Section 5: Modeling**

5a. Detailed descriptions of all models trained, including:

* All features used in the model (including any feature selection).
* Types of models implemented (e.g., Random Forests, Neural Networks).
* Loss function used.
* Coefficients where possible (e.g., linear regression or decision trees with a small number of features).

5b. Justification for the choice of model types implemented.

5c. Method for evaluating the model(s) reported in the paper, including details of train-test splits or cross-validation folds.

5d. Method for selecting the model(s) reported in the paper.

5e. For the model(s) reported in the paper, specify details about the hyperparameter tuning:

* Range of hyper-parameters used and a justification for why this range is reasonable.
* Method to select the best hyper-parameter configuration.
* Specification of all hyper-parameters used to generate results reported in the paper.

5f. Justification that model comparisons are against appropriate baselines.

**Section 6: Data leakage**

6a. Justification that pre-processing (Section 4) and modeling (Section 5) steps only use information from the training dataset (and not the test dataset).

6b. Methods to address dependencies or duplicates between the training and test datasets (e.g. different samples from the same patients are kept in the same dataset partition).

6c. Justification that each feature or input used in the model is legitimate for the task at hand and does not lead to leakage.

**Section 7: Metrics and uncertainty**

7a. All metrics used to assess and compare model performance (e.g., accuracy, AUROC etc.). Justify that the metric used to select the final model is suitable for the task.

7b. Uncertainty estimates (e.g., confidence intervals, standard deviations), details of how these are calculated, and a justification for why they are valid for each metric reported in the paper.

7c. Justification for the choice of statistical tests (if used) and a check for the assumptions of the statistical test.

**Section 8: Generalizability and limitations**

8a. Threats to external validity.

8b. Contexts in which the authors don’t expect the study’s findings to hold.

1. Note that this is a high bar for computational reproducibility. It might not be possible to provide such a script—for instance, if the analysis is run on an academic computing cluster, or if the dataset does not allow for programmatic download. [↑](#footnote-ref-0)