Instructions for

Environmental Machine-learning, Baseline Reporting, And Comprehensive Evaluation: The EMBRACE Checklists

General Introduction

Checklist: Version 1.0. Last Updated Date: September, 2024.

Instructions: Last Updated Date: October, 2024.

This instructions document provides additional explanations for the Environmental Machine-learning, Baseline Reporting, And Comprehensive Evaluation (EMBRACE) Checklist. The main purpose of the checklist is to facilitate the minimum reporting and comprehensive evaluation for data and methods used for machine learning (ML) research in environmental science and engineering (ESE). We warmly encourage researchers to include the checklist as a reference, and if interested list it as supplementary materials in their publications.

In addition to this instructions, users could learn additional information from the following publications:

- A **Review Article**. Zhu, J.-J., Yang, M., & Ren, Z. J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environmental Science & Technology*, 2023, 57(46), 17671-17689. https://doi.org/10.1021/acs.est.3c00026.
- A **Viewpoint Article**. Zhu, J.-J., Boehm, A. B., Ren, Z. J. EMBRACE Checklist: Environmental Machine-learning, Baseline Reporting, And Comprehensive Evaluation. *Environmental Science & Technology*, 2024, https://doi.org/10.1021/acs.est.4c09611.

Interactions

- All relevant materials and resources can be found and downloaded from the author’s GitHub repository: https://github.com/starfriend10/EMBRACE. The repository is also served as a community-owned platform to share data and codes, refine forms/checklists, discuss ML-ESE topics, promote research outcomes, and continuously identify/avoid pitfalls and follow good practices.
- Future updates, development, and new resources can also be found in the same GitHub repository.
- We encourage you to share your checklist. One approach is to include it as supporting information along with your publication. Alternatively, we also welcome researchers to share their checklists via the repository. If you’re willing to do so, please find detailed steps provided in the GitHub repository where we will help verify, upload, and share the document accordingly.
- Filled checklist can immediately be shared (recommended), and you can also save it as a read-only document. If you use Microsoft Windows, please follow these step: “File” → “Print” → Select “Microsoft print to PDF” in “Printer” → Print → Save it as a new PDF document. If you use macOS, there will be additional steps (e.g., AirPrint in the Protocol) needed to “print to PDF”, please see the instructions at the repository for details.
- Please cite the Viewpoint when using the form or checklist. If you have any questions or suggestions, please send an email to Dr. Jun-Jie Zhu at Princeton University (junjiez@princeton.edu or ranmuweijie@gmail.com).

Document History

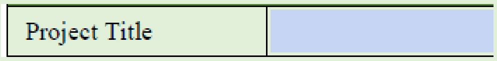
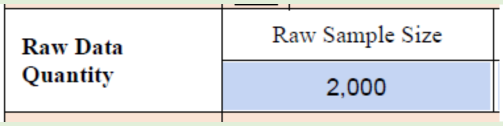
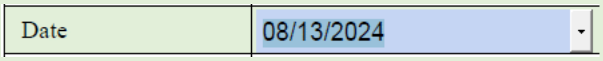
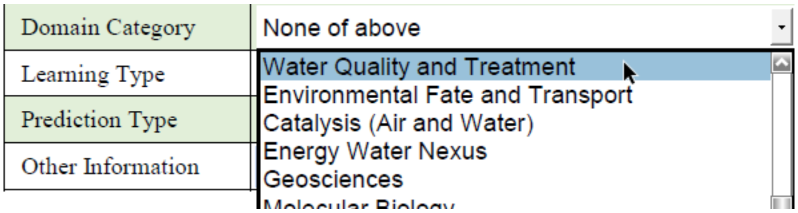
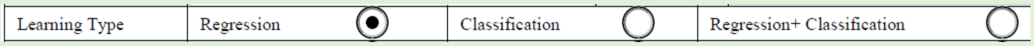
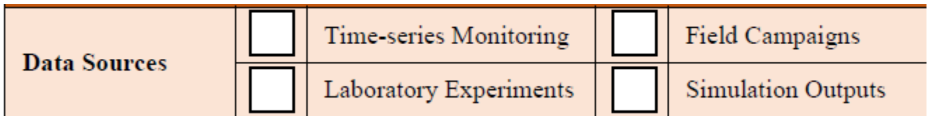
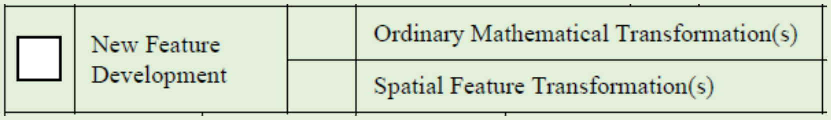
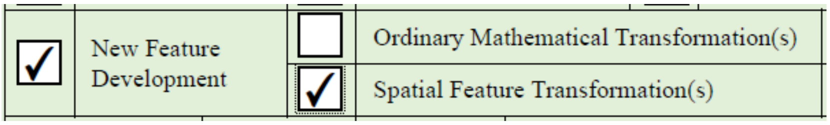
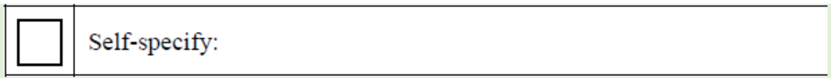
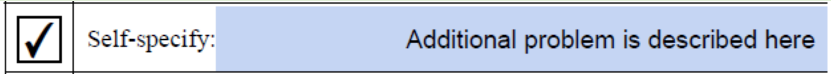
08/2023	Project was initialized
09/2024	Version 1.0 was published

Terms and Acronyms Used in the Checklists and Explanations

Items	Additional Explanations
Data size	The number of data units, e.g., sample size × variable size if there are no missing values
Sample size	The number of samples or observations
Variable size	The number of initial variables, e.g., environmental factors
Feature size	The number of predictors for a model, which can be less than or more than variable size
SFR	The ratio of sample size to feature size
Training-Testing (TT)	Data splitting to training (seen) and testing (unseen) datasets
(Pre-)Training-Validation-Testing (TVT)	Data splitting to training (seen) and testing (unseen) datasets, and the training dataset is divided into pre-training and validation subsets.
Cross-Validation-Testing (CVT)	Data splitting to training (seen) and testing (unseen) datasets, and the training dataset is applied with cross-validation (CV).
Testing ratio	The ratio of testing sample size to all sample size
Parameter	Internal constants, coefficients, or weights that are learned or estimated purely from data
Hyperparameter	External model configurations that facilitate a better determination of model parameters
Statistical method/model	There is no clear boundary between statistical method and ML method, here we define them to better differentiate their usages. The methods/models use less or no hyperparameters, tend to fit the model parameters (e.g., coefficients) from data, and commonly have higher interpretability
ML method/model	There is no clear boundary between statistical method and ML method, here we define them to better differentiate their usages. The methods/models that strongly rely on hyperparameters (e.g., number of neurons or trees), commonly have a more explicit, iterative learning process (e.g., epoch (an iteration of training process for a neural network)), and/or have lower interpretability
Statistical data analysis	Conventional data analysis using statistical approaches other than regression and classification, e.g., descriptive statistics, *t*-test, and ANOVA
Grouped random	Data splitting approach that bundles groups of data into the same subset
Simple block	Data splitting approach that simply separates the data into subsets based on the sequence
Sub-group random	Data splitting approach that bundles relevant data points together and randomly assign them to same subset
Sub-group block	Data splitting approach that bundles relevant data points together and assign them to same subset based on the sequence
Data leakage	Information of unavailable or unseen data (raw or transformed values, distribution, etc.) is included in the model training process, leading to biased results

Majority of the above explanations are quoted from Zhu et al. (2023).

Format and Types of Information Filled in the Checklist

Type	Example
Text	Project Title:
Number	**Raw Data Quantity** Raw Sample Size: 2,000
Date	Date: 08/13/2024
Dropdown list	Domain Category: None of above Learning Type Prediction Type Other Information Water Quality and Treatment Environmental Fate and Transport Catalysis (Air and Water) Energy Water Nexus Geosciences Molecular Biology
Single Choice	Learning Type: Regression ● Classification ○ Regression+ Classification ○
Multiple Checkmarks	**Data Sources** □ Time-series Monitoring □ Field Campaigns □ Laboratory Experiments □ Simulation Outputs
Conditional Checkmarks	*Before* □ New Feature Development Ordinary Mathematical Transformation(s) Spatial Feature Transformation(s)
	After ☑ New Feature Development □ Ordinary Mathematical Transformation(s) ☑ Spatial Feature Transformation(s)
Conditional Text	*Before* □ Self-specify:
	After ☑ Self-specify: Additional problem is described here

The checklist facilitates the reporting of minimal information and comprehensive step-by-step self-check used for ML research in ESE. We strongly recommend you to use it as supporting information when submitting your research manuscript. If your study includes multiple datasets or multiple objectives, use the checklist for each dataset or objective, or select a representative case. For example, a study is to predict arsenic (As) and lead (Pb) concentrations in groundwater based on two datasets and two ML models. If arsenic is the primary target contaminant, the user can prepare a checklist that focuses on arsenic concentration prediction.

The table of "**Project Overview**" helps to record the general information of your ongoing or finished study. You can also use it to track potential problems during your research. The checklist is specifically designed for environmental and sustainability researchers; however, it can also be used for all ML researchers. You can specify your domain area if your research topic goes beyond the defined domain categories. This checklist will also be served as additional data for your research publication. If your work is not published yet, you are also welcome to provide relevant information (e.g., as preprint) so your work may potentially be accessed by more readers.

I. Problem Formulation

You are invited to provide project objective(s), helping differentiate different studies in a similar subarea. The feasibility assessment should be conducted from 7 aspects of your study. While there are no quantitative references being used for your choices because they all vary for different studies, you may need to acquire additional information to complete this self-evaluation based on your best knowledge. The overall feasibility is based on the needs of resources for your research, and the weights of the seven aspects could be different in many cases. For example, if your working problem is a relative simple classification without the need of high computation resources, a low level of computation resources will not be the major barrier for obtaining reasonable outcomes. Check additional explanations below to weigh the level of each aspect (low, medium, and high) in your study.

Items	**Additional Explanations (example factors to be considered)**
Data Availability	Will you have sufficient sample size and variable diversity? Will the data collection involve retrieval difficulty and requirements for permission or similar data ethics considerations? Compared to typical ML research work, will you have higher or lower sample size and variable diversity? How about when comparing with other similar studies?
Model Accessibility	Will you have readily accessible software or platform for modeling? Will the accessible package or utilities provide necessary and sufficient capabilities for your modeling work? Will you develop the model from starch that requires additional resources?
Computation Resources	Will your modeling require local or cloud computing? Will your computation need CPU or GPU resources? Will your data collection, preprocessing, and input/output require extend storage needs? Have been these computation resources secured?
Knowledge Preparation	Do you have sufficient ML knowledge to develop the models without basis? Do you have sufficient (environmental) domain knowledge to better prepare data (e.g., feature engineering), understand model (e.g., model explanation), and interprets results (e.g., causality analysis)?
Time Availability	Will you have sufficient time to conduct the study even if additional time is needed for correction, uncertainty analysis, and post-hoc analysis?
Financial Availability	Will you have sufficient financial support to fulfill the above needs? Will you have additional financial backup for unexpected needs?
Risk Tolerance	To what extent are you prepared for the study to not yield satisfactory results? Are you flexible that if the focus of the study has to be shifted from ML to conventional statistical modeling or only comprehensive statistical analysis?

II. Data Collection

This table includes typical information needs to be filled associated with data collection. Multiple checkboxes can be checked for the same topic (e.g., a study can have multiple data sources). Additional explanations are listed below.

Items	Additional Explanations
Data Sources	• **Field campaigns** are those activities to collect field data which may need additional measurements in the lab, different from **laboratory experiments** that the main experiment and measurement are both conducted in the lab.
Source Types	• This item is mainly to help design a better data splitting method. • **Time-series** (or sequential) data are only when the data have fixed (time) intervals, including when irregular intervals preprocessed to regular intervals. It is possible to have **time-series** data obtained from journal/conference publications (**literature-based**). **Ordinary** means no apparently time-series sequence without using literature-based data. • This part will link with the section VII. Data Leakage Management
Data Types	• Numerical data can be categorical depending on the raw format (e.g., not continuous) and feature engineering. • Visual data include images, videos, etc., graphical data include graphs of molecules, chemicals, etc., and auditory data include sounds, voice, etc.
Data Ethics	• **Special ethical considerations** can be important for experiments such as those related to animal or human, or those data related to personal information, etc. • **Special permissions** are often required to obtain and use data under copy-right or other intellectual property rights. • For data sources that are freely retrieved and used, **giving appropriate credits** is essential.
Data Availability	• We encourage to share data along with the publication (DOI), or through a PURL (e.g., Github). • If the data are not available at the time of publication, but will be released in the future or available as requested, it will be great to include such information in SI.
Raw Data Quantity	• **Raw sample size** is the number of observations in the initial dataset. While the need of sample size varies based on objective and method, there is a minimal need of samples to develop good models. Bigger sample size could also allow a margin to remove noises or lower quality data. • **Raw variable size** is the number of variables (e.g., environmental factors) in the initial dataset. • **Raw total data volume** is the number of total valid data where missing data should be excluded, e.g., Raw sample size × Raw variable size – Missing data. • **Raw sample-variable ratio** is Raw sample size/Raw variable size, only indicates a general information density and initial assessment of data availability. A minimal 10 is needed for classification, and minimal 100 is recommended for regression.
Raw Data Quality	• Quality assurance and quality control (**QA/QC**) improves the data quality and is typically required for data report to government agency (e.g., USEPA). • Data retrieved from **peer-reviewed sources** may likely have higher quality

Examples of Common Problems

- Basic information about data volume and variables are not clearly provided in the publication.
- Low raw sample size or low raw sample-variable ratio.
- Utilize private data or materials with copy-right without obtaining appropriate permissions.

Examples of Less Robust Practices

- A study is to classify water quality levels based on 60 raw samples, while the underlying relationship is complex.
- A study is to predict air pollutant based on 1000 raw samples but with 200 raw variables (unless a good feature reduction will be performed to reduce the redundant variable size).
- A study is to predict soil contaminant concentrations based on 500 raw samples with 50 raw variables, but half of the variables have missing data% > 30% (unless these variables occupy missing values at the same soil samples).

III. Data Preprocessing

This table lists typical data preprocessing procedures and methods. Checking a checkbox means that you carefully considered or applied it. For example, your dataset doesn't have missing values, but you should check the box of missing data management if a verification of data completeness is conducted. Additional explanations are listed below.

Items	Additional Explanations
Data Cleaning	• **Abnormal data management** takes care of noisy data such as garbled or duplicates, etc. • **Statistical outliers** are detected based on statistical distribution, and **technical outliers** are detected based on domain knowledge. Check the option after applying the practice, even with a complete valid dataset and no outliers. Reporting relevant steps is also important. Technical outliers may stem from measurement errors, detection limits, or system maintenance. **Data correction** may be applied with a reasonable assumption for the values.
Data Enrichment	• **Missing data management** include deletion, replacement, interpolation, imputation, etc. • Check the option after applying the practice even with a complete dataset. Reporting relevant steps is also important. • **Data augmentation** is frequently used to visual data (e.g., image rotation, random cropping, etc.) but can also be applied to numerical data (e.g., adding noises, smoothing). • **Data aggregation** converts high-resolution data to low (e.g., hourly to daily) in temporal, spatial, or other dimensions. • **Group ratio** can be important when dealing with classification problem with imbalanced data as it can generate bias in evaluation. Please provide the maximum ratio of sample size among different groups (**Upper sample ratio among groups**) if classification is studied. In this case, **data balancing** (e.g., upsampling/downsampling for minor/major classes, respectively) is strongly recommended.
Feature Engineering	• **Initial variable exclusion** includes factors like low data quantity/quality and domain knowledge. • In new feature development, **categorical transformations** include one-hot and ordinal encoding. • Excluding a dummy feature if one-hot encoding is used. • **Ordinal encoding** is one common way to transform features, however, it is essential to understand that ordinal number should correctly reflect the characteristics of features. • **Mathematical transformations** create new features using mathematical functions. • **Time-dependent transformations** generate new temporal features (e.g., t-3, t-2, t-1) based on historical data and periodic patterns. • **Spatial transformations** create new features based on neighboring features. • **Feature extraction** identifies important features in textual and image data. • **Feature selection** or dimensionality reduction (e.g., PCA) selects important features or reduces feature numbers. Feature importance analysis ranks feature contributions. • **Feature scaling** (for Y) can be crucial, especially when Y spans several orders of magnitude. • **Circular features** may use trigonometric or other transformation methods to avoid incorrect representation. • **Sparse feature dataset** should be preprocessed to gain higher information density.
Data Splitting	• Please refer to page 2 or Zhu et al. (2023) for terminologies and concepts. Sub-Group R/B means Sub-Group Random or Block splitting • When collecting data, it is important to retrieve data from different scenarios. It helps to build a dataset with high variable diversity and high information density.
Final Data Quantity	• The testing ratio is the proportion of the testing dataset to the entire dataset. Ratios that are too low or too high are unfavorable as they may be less representative or result in too small training data for model learning. A range of 0.1-0.4 is recommended. • "Features" are the final variables after preprocessing, just before model training. A higher SFR indicates better learning materials with higher information density. A minimum of 10 is needed for classification problems, while 100 is recommended for regression problems.

Examples of Common Problems

- A complete, valid dataset is used with a quality check but this information is not described in the publication.
- Insufficient data cleaning fails to exclude low-quality data, further resulting in less reliable models.
- Feature selection or reduction is not applied to a dataset with a large number of sparse features.
- Final SFR is less than 10.
- No validation dataset is prepared.

Examples of Less Robust Practices

- A study is to classify landslide risks using geophysical and meteorological conditions but based on a serious. imbalanced data (10% of data represent high landslide risk) and an ordinary evaluation metric (e.g., RMSE) is used.
- A study collects 500 samples and develops a dataset based on 500 molecular fingerprint-based features to predict chemical property. No feature reduction is applied.

IV. Methods Selection and Model Initialization

This section documents selection of methods and model initialization. Additional explanations are listed below.

Items	Additional Explanations
Methods Selection	• **Selected methods based on case needs**: The selection of methods is highly recommended based on resources available, such as data size or computational resource. • **Number of supervised learning methods** counts both statistical and ML methods. When counting the methods, it is recommended to consider the methods originated from different root methods. For example, autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) should be considered as the same root method, while random forest and extra trees are similar but two different methods. Furthermore, the number of methods is counted only for the core supervised learning methods. For example, assuming that two feature selection methods are used to prepare two datasets to develop two respective random forest (RF) models, only one ML method is counted in this case. If multiple ML methods are studied but no statistical method is involved, set M = 0. Check page 2 for additional explanations to differentiate ML and statistical methods. • **Deep ML** commonly includes multiple hidden layers, or complex structures, such as convolutional neural networks (CNNs) and transformer. • **ML family** includes tree-based (e.g., RF), neural network (e.g., recurrent neural networks), or others (e.g., support vector regression). • Please provide the **Final Optimal Method(s)** after methods/models comparison.
Model Uncertainty	• **Examine random seeds**: Repeating the experiment based on the same modeling structure but different external settings to understand variability in performance. The examination should cover a certain range of seeds without an intentional selection. • **Replication of modeling process**: Repeating the experiment based on the same modeling structure and external settings to understand variability in parameter randomness. This can be part (e.g. repeating training or CV) or whole modeling processes.

Examples of Common Problems

- Only a single method is tested and applied without thoroughly justifying the approach and describing limitations.
- Advanced ML methods are studied, but no statistical method is included without justification or description of limitation.
- The tested methods demand larger resources that are not readily allocated.

Examples of Less Robust Practices

- A study plans to use 200 samples to develop a neural network model (> a few thousand samples are recommended).
- A random forest model outperformed a neural network model based on 200 samples in a previous study. After collecting more samples to 1000, a following study is conducted based on only the random forest method.
- A study is to initialize a deep learning investigation on predicting a water treatment process performance, but computational resources (e.g., multi-core CPUs or GPUs) are not appropriate secured.

V. Model Evaluation and Model Optimization

This table documents reporting of methods in model evaluation and model optimization. Additional explanations are listed below.

Items	Additional Explanations
Model Evaluation	• Set a **goal model performance** to help navigate the modeling iteration process. • Report the used evaluation metrics: Ideally two complementary metrics are used. For example, R^2 and RMSE which are complementary metrics based on different underlying quantification strategies, while R^2 and adjusted R^2 are metrics from the same root. • Efforts to check for **Under-/Over-fitting** helps to mitigate model bias. Final comparison should be only based on unseen, testing dataset. If imbalanced data are used, either pretreating the data to reduce the imbalance, or employ **weighted metrics** to account the bias. • ML models can significantly outperform statistical models, but in many cases statistical models could perform close, similar, or even better than ML models.
Hyper-parameter Optimization	• **Metaheuristics** include various automatic, intelligent searching methods. Typical examples include genetic algorithm, particle swarm optimization, Bayesian optimization, etc. • **Number of supervised learning method(s) tuned by HPO** may only count for ML methods. • **HPO**: It will be meaningful to tune at least two hyperparameters and three options for each hyperparameters. Metaheuristics commonly search a continuous space of hyperparameter values. **Fair comparison**: Other than statistical methods, ML methods should be all optimized before comparison. • **Pre-defined CV**: This is needed particularly when the data is uneven distributed, such as literature-sourced data. To avoid data leakage issues, pre-defining splits before CV is useful.

Examples of Common Problems

- Final evaluation is based on only one metric (e.g., RMSE only).
- Overfitting is observed but no further intervention (e.g., regularization, early stopping, and dropout) is applied, and limitations of approach are not considered.
- HPO is not applied to search optimal models.
- When calculating R^2 as one of the model performance metrics, using the fitting equation $y=ax+b$ based on real data and predicted values. (The correct way is to calculate R^2 based on the fitting equation y=x)

Examples of Less Robust Practices

- A study develops a novel ML model with appropriate HPO and reports that the model outperforms a random forest (RF) baseline model. However, the RF model had not been optimized.
- A study applies HPO using expert experience, but the testing dataset shows much higher accuracy compared to the training dataset. This means that the model is underfitting, so a better HPO is recommended.
- A study collects 200 samples and 60% of the data (120 samples) are used to train a model. A 10-fold CV is applied and each validation set only has 12 samples to be evaluated, generally considered too small to be reliable.

VI. Model Explanation

This table focuses on model explanation where additional analysis and domain knowledge are required to better understand the model characteristics and behaviors. Additional explanations are listed below.

Items	Additional Explanations
Model Interpretability	• More information about model interpretability can be found in many sources, such as Amazon Web Services [External link 1] and [External link 2].
Model Explainability	• The **limitation** of model explanation comes from both the limitation of model explanation methods (e.g., SHAP) and the intrinsic complexity of ML models. • **Feature importance analysis** (FIA) can be performed during feature selection, which is not part of model explanation results. • Model explanation should be performed consistently using the **final/optimal model** that is used for prediction. “Final model” is the model that outperforms other models and is selected to be used for prediction and explanation in the final. • **Model explanation is not always correct**, it only reflects the model inner working mechanisms, which not necessarily represent the relationships between inputs and output(s) in real world. • Prediction accuracy is not necessarily directly associated with model explanation.
Causality	• Higher prediction accuracy may represent a better construction of the relationships between inputs and output(s), which **not necessarily represents** more logical causal relationships. • As mentioned above, model explanation only reflects the model inner working mechanisms, which **not necessarily represent** the relationships between inputs and output(s) in real world. • FIA results may have **illogical or counterintuitive relationships** (positive/negative or magnitude) based on domain knowledge, it will be a good practice to 1) identify and 2) study these unique but useful patterns.

Examples of Common Problems

- A study developed a model based on a complex, advanced ML method, but did not provide appropriate model explanation.
- Environmental domain knowledge was not applied to align/verify the model explanation results.
- Simply believe that high accurate models always give correct casual relationships without analyzing feature relationships.

Examples of Less Robust Practices

- A study reports that a complex ML model performs similarly or slightly better than a statistical model but still uses the ML model as the final model to predict and understand the relationship among environmental factors. (Note: Statistical models typically have a much higher model interpretability. Understanding casual relationships between environmental factors is equally important as a good model accuracy. In this case, the statistical model is recommended to use for the high interpretability and good model performance).
- A study obtains a model with a high accuracy, but the model explanation exhibits a clear counterintuitive relationship between variables based on domain knowledge. No further investigation is conducted.
- A study investigates several ML methods, and applies one ML model to predict target variable and utilizes another ML model to explain the feature relationships.

VII. Data Leakage Management

This table summarizes possible data leakage issues based on the author's experiences and knowledge. While there could have other potential problems due to various scenarios and use cases, users can specify them in the bottom of the table. More information about splitting methods can be found in Zhu et al. (2023) Figures 3 and 4. Additional explanations are listed below.

Items	Additional Explanations
General/Data Source	• **Verified no leakage from response variable Y**: While it is beneficial to include features that closely correlated with Y, it is important to ensure that these features are independent from the response variable Y. • **Verified no leakage from *future* to *current***: For example, data leakage issues stemmed from temporal preparation or preprocessing which may include impossible "future" data to predict "current data". A straightforward way is to conduct a "thought experiment" to verify each component required if their data are ready to be used onsite and if the model could work as planned (e.g., model input data acquisition is feasible in terms of time and accuracy)? • **Verified no leakage from *current* to *current***: For example, data leakage issues stemmed from temporal or spatial preparation or preprocessing which may include impossible "current" data to predict "current data". The information should not be readily available due to temporal or spatial constraints (e.g., time or distance). The above approach may also be used to verify. • More information about **grouped random splitting** can be found in Zhu et al. (2023) Figures 3 and 4. Similarly, information about other splitting methods is also provided.
Data Enrichment	• If no missing data management is needed, select not available (N/A). • Replacing missing data with a representative value (e.g., mean or median of the feature data) is a common practice, which should be only based on the training data when CV is not applied. Similar manipulation is applied for the model-based imputation.
Feature Engineering	• If no such practice is needed, select not available (N/A). • Similar as above, feature engineering or other preprocessing should be only based on the training data when CV is not applied, and then apply the training-based transformation to the testing dataset.
CV Loop	• When CV is applied, strictly, all data processing and feature engineering should be performed within each split (i.e., the method is the same but the values can vary for each split).

Examples of Common Problems

- Data leakage issues are not considered or not well-managed.
- The random splitting is used without considering data sources and types
- Feature engineering or feature scaling is developed based on the entire dataset when TVT is used
- Feature engineering or feature scaling is developed based on the entire training dataset when CVT is used

Examples of Less Robust Practices

- A study aims to predict current day's (*t*) BOD_5 concentrations in wastewater based on lab measurement data, but uses recent four days' (*t-1, t-2, t-3,* and *t-4*) BOD_5 concentrations as predictors. This is a case of "leakage from future to current" because recent four days' BOD_5 concentrations are future information and not available at the current day.
- A study is to predict resource recovery rate based on 1000 samples collected from literature, the grouped random splitting is not appropriate applied. This would result in over-optimistic results, check Yang et al. (2023, 2024) for examples.
- A study collects data representing spatial distribution across a large region and a small fraction of missing data are filled by interpolation using neighbor values. All data are then random splitting into training and testing subsets. This would also potentially lead to over-optimistic results as the training and testing utilize the same information.

VIII. Additional Items

If possible, users may also contribute to this table by providing description and example of new items when the items are found to be representative and important. Additional explanations are listed below.

Items	Additional Explanations
Miscellaneous	• We encourage include links to data and open source code used for the target research, especially code that uses novel methods or frameworks. • Data and code may be available as requested, or restricted under a limited license.
Self-specify	• You can add any additional information about any of the sections here. • Some exceptional examples can be specified here. For example, you have a small dataset but you're using transfer learning to enhance the model performance.

References

Zhu, J.-J., Boehm, A. B., Ren, Z. J. EMBRACE Checklist: Environmental Machine-learning, Baseline Reporting, And Comprehensive Evaluation. *Environmental Science & Technology*, 2024. https://doi.org/10.1021/acs.est.4c09611

Zhu, J.-J., Yang, M., & Ren, Z. J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environmental Science & Technology*, 2023, 57(46), 17671-17689. https://doi.org/10.1021/acs.est.3c00026

Yang, M., Zhu, J. J., McGaughey, A., Zheng, S., Priestley, R. D., & Ren, Z. J. Predicting extraction selectivity of acetic acid in pervaporation by machine learning models with data leakage management. *Environmental Science & Technology*, 2023, 57(14), 5934-5946. https://doi.org/10.1021/acs.est.2c06382

Yang, M., Zhu, J.-J., McGaughey, A. L., Priestley, R. D., Hoek, E. M., Jassby, D., & Ren, Z. J. Machine Learning for Polymer Design to Enhance Pervaporation-Based Organic Recovery. *Environmental Science & Technology*, 2024, 58(23), 10128–10139. https://doi.org/10.1021/acs.est.4c00060

Other useful links

EMBRACE Checklist GitHub Repository: https://github.com/starfriend10/EMBRACE.

Model Interpretability and Explainability: https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html (accessed on 10/20/2024)