The EMBRACE Checklist

Version 1.0. Last Updated Date: September, 2024.

The main purpose of EMBRACE Checklist is to facilitate the reporting of basic data and methods used for supervised machine learning (ML) research in environmental science and engineering (ESE). It also aims to provide a comprehensive evaluation of important preparation and analysis steps. We warmly encourage researchers to include the checklist as a reference, and if interested list it as supplementary materials in their publications.

Detailed instructions and additional resources are available open access for download from the author's GitHub: https://github.com/starfriend10/EMBRACE

- Future updates and new resources will be available in the above GitHub repository, which 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.
- We encourage you to direct share your checklists, but you can also save it as a read-only document. See instructions for detailed steps.
- Please cite the Viewpoint when using the checklist. If you have any questions or suggestions, please contact Dr. Jun-Jie Zhu at Princeton University (junjiez@princeton.edu or ranmuweijie@gmail.com).

Project Overview					
Project Title					
Contributing Authors					
Date		Completed by			
Contact Email		DOI (if published)			
Domain Category	None of above	Or Specify:			
Learning Type	Regression ◉	Classification ○	Regression+ Classification ○		
Prediction Type	Deterministic ◉	Probabilistic ○	Deterministic+ Probabilistic ○		
Other Information					

I. Problem Formulation							
Project Objective(s)							
Feasibility Assessment	Data Availability	Model Accessibility	Computation Resources	Knowledge Preparation	Time Availability	Financial Availability	Risk Tolerance
Levels	Low	Low	Low	Low	Low	Low	Low

II. Data Collection								
Data Sources	☐	Time-series Monitoring	☐	Field Campaigns	☐	Government Database	☐	Scientific Literature
	☐	Laboratory Experiments	☐	Simulation Outputs	☐	Others. Specify:		
Source Types	☑	Ordinary	☐	Time-series	☐	Literature-based	☐	Others. Specify:
Data Types	☐	Continuous Numerical	☐	Categorical	☐	Textual	☐	Visual
	☐	Graphical	☐	Auditory	☐	Others. Specify:		
Data Ethics	☐	No special ethical considerations are required, or followed the necessary requirements when needed						
	☐	No special permissions are required, or obtained the necessary permissions when needed.						
	☐	Provided appropriate credit(s) to the data source(s)						
Data Availability	☐	Data are publicly available through a PURL or DOI			☐	Data are available as requested		
Raw Data Quantity	Raw Sample Size		Raw Variable Size		Raw Total Data volume		Raw Sample-Variable Ratio	
Raw Data Quality	The raw data had an internal QA/QC before data retrieval?				Yes ○		No ◉	
	The raw data went through a peer-review or similar process?				Yes ○		No ◉	

III. Data Preprocessing								
Data Cleaning	☐	Abnormal Data Management	☐	Statistical Outlier Management	☐	Technical Outlier Management		
	☐	Data Correction due to Technical Limitations	☐	Others. Specify:				
Data Enrichment	☐	Missing Data Management	☐	Data Augmentation	☐	Data Aggregation	☐	Data Balancing
	Upper sample ratio among groups:		☐	Others. Specify:				
Feature Engineering	☐	Initial Variable(s) Exclusion	☐	Feature Importance Analysis	☐	Feature Extraction		
	☐	Categorical Feature Transformation	☐	Dummy Feature Removal				
	☐	Ordinal Encoding	Encoded variable(s):					
	☐	Feature Selection	☐	Feature Scaling (for X)	☐	Feature Scaling (for Y)	☐	Circular Feature Scaling
	☐	New Feature Development		Ordinary Mathematical Transformation(s)		Time-Dependent Feature Transformation(s)		
				Spatial Feature Transformation(s)		Others. Specify:		
	Sparse Dataset?	Yes ○	No ◉	Preprocessed to increase information density?	Yes ○	No ◉		
	☐	Normality Test (before/after transformation)	☐	Others. Specify:				
Data Splitting	Framework	◉	Training-Testing	○	(Pre-)Training-Validation-Testing	○	Cross-Validation-Testing	
	Splitting Method	☐	Simple Random	☐	Grouped Random	☐	Simple Block	☐ Sub-Group R/B
	☐	Ensured training dataset covered most possible scenarios (e.g., different seasons)						
Final Data Quantity	Testing Ratio	Preprocessed Sample Size	Preprocessed Feature Size	Sample-Feature Ratio	Training SFR			

IV. Methods Selection and Model Initialization							
Methods Selection	☐	Selected supervised learning methods to be tested based on case needs (e.g., data, computation resources, etc.)?					
	Methods Studied: N ML and M Statistical			N =		M =	
	ML Family	☐	Tree-based	☐	Neural Network	☐	Others
	ML Type	☐	Shallow ML	☐	Deep ML	Final Optimal Method(s):	
Model Uncertainty	☐	Examined Random Seeds	☐	Replicated Modeling Process			
	☐	Assessed probabilistic prediction results	☐	Others. Specify:			

V. Model Evaluation and Model Optimization								
Model Evaluation	☐	Set a goal model performance before evaluation						
	Specify Evaluation Metrics:							
	☐	Studied under-fitting issue	☐	Studied over-fitting issue				
	☐	Final evaluation was based on the testing data	☐	Weighted metrics were used for imbalanced data				
	☐	ML significantly outperformed statistical models	☐	Statistical perform similar or even better than ML models				
Hyper-parameter Optimization	HPO Method	☐	No HPO	☐	Trial-and-Error/Experience	☐	Grid/Random Search (similar)	☐ Metaheuristics
		☐	Other HPO. Specify:					
	Number of methods tuned by HPO:			Number of hyperparameters searched:				
	Number of values in each hyperparameter:		○	Continuous Search	◉	Discrete Search. Specify the number:		
	CV Method	◉	No CV	○	k-fold CV	○	Leave-one-out CV	○ Stratified CV
		○	Pre-defined or other CV. Specify:					

VI. Model Explanation				
Model Interpret-ability	☐	Reported model interpretability as intrinsic property	☐	Described tradeoff between interpretability and complexity
	☐	Explained limitations of low model interpretability	☐	Compared the methods with different interpretabilities
Model Explain-ability	☐	Reported the explanation method(s) used for FIA	Explanation method(s):	
	☐	Reminded high accuracy might not be trustable explanation	☐	Described application of explanation on the optimal model
	Reflected on understanding that explanations:	☐	couldn't cover all mechanisms	☐ might be incorrect
Causality	☐	Reflected on higher accuracy not implying logical causal relationships		
	☐	Reported understanding that explanation results do not indicate causality		
	☐	Described alignment of explanations with environmental domain knowledge for causality		
	☐	Identified illogical or counterintuitive parts based on environmental domain knowledge		
	☐	Described study of illogical or counterintuitive parts based on environmental domain knowledge		

VII. Data Leakage Management				
General/ Data Source	Verified no: ☐ response variable Y leakage	☐ future-to-current data leakage	☐ current-to-current data leakage	
	Confirmed that no overlapped data between training and testing	Yes ○	No ◉	
	Data splitting method for literature-source data:			
Data Enrichment	Verified that missing data replacement utilized only seen dataset	N/A ○	Yes ○	No ◉
	Verified that missing data interpolation utilized neighbor data in same subset	N/A ○	Yes ○	No ◉
	Verified that model-based imputation utilized only seen data	N/A ○	Yes ○	No ◉
Feature Engineering	Confirmed that feature engineering utilized only seen dataset	N/A ○	Yes ○	No ◉
	Confirmed that feature scaling utilized only seen dataset	N/A ○	Yes ○	No ◉
	Confirmed that data splitting method for time-dependent data:			
CV Loop	Verified that missing data replacement utilized only seen data in each split	N/A ○	Yes ○	No ◉
	Verified that missing data interpolation with neighbor data in each split	N/A ○	Yes ○	No ◉
	Verified that feature engineering utilized only seen data in each split	N/A ○	Yes ○	No ◉
	Verified that feature scaling utilized only seen data in each split	N/A ○	Yes ○	No ◉
	Data splitting method for literature-source data in CV loop:			
	Data splitting method for time-dependent data in CV loop:			
Others	☐	Self-specify:		

VIII. Additional Items		
☐	PURL or DOI are publicly available. Data:	Code:
☐	Email or URL are available to request. Data:	Code:
☐	Self-specified Item:	