

Data analysis with ML – Draft Guidelines

Richard Tyson – Pierre Chatagnon

CLAS Collaboration Meeting - 17th of November 2026

Why ?

- Concerns were raised at previous Collaboration meetings on how the review of analyses with ML should proceed.
- In particular, there were discussions on whether we need input from AI/ML experts.
- **A survey circulated to reviewers with analysis including ML, in September 2025**, seems to indicate the latter is not necessary even if reviews were more challenging than normal
- The two main problems seem to be:
  - a lack of relevant information in analysis notes that slows down the review process.
  - concerns about being able to identify bias introduced in physics results
- These guidelines aim to prevent this by spelling out what should be in an analysis note and what to look for to prevent bias in physics results
- It also a give short intro to ML and examples as a starting point for non-ML expert reviewers.

Question 2: Did you find this review particularly challenging compared to others without AI/ML components?

- Somewhat yes, for 4 reasons at least:
  - When using ML technics, the analysis author sometimes omits to describe the 'classical' part of their analysis, which induces delays
  - ML is sometimes used to hide problems, and rarely compared with 'classical analysis' when relevant
  - The validation of the ML technics/algorithm is sometimes neglected (see previous point)
  - ML technics are sometimes used in region where they have not been trained, and a specific validation for this should be done
- I'm not an expert in AI/ML, therefore I'm not able to judge the details of the various techniques implemented. I mostly focused on reviewing how they are implemented and applied to get the results.
- If the ML method is the main aspect of the study, other focuses have to be set for the review than for a classical analysis with simple cuts. Especially more details and plots are needed to really verify the validity of the procedure.
- Yes, somewhat. The challenge stemmed not from the BDT method itself, but from the way it was applied in a kinematic region not covered by existing data, raising concerns about extrapolation and validation.

Question 4: Did you have trouble with understanding:

Question 4c: The impact of the method on the results:

- No.
- See discussion with comparison with traditional methods.
- It needs to be validated, if the ML method produces correct results or fake results (i.e. with a modulation but not a rejection of backgrounds) which only look correct.
- Yes, due to insufficient validation in the extrapolated region.

Question 4d: How to validate the method:

- No.
- This is really the heart of the question. See discussion on MC comparison above
- We requested additional validation checks, like implementing modulation in the data and the background MC and trying to reconstruct them.
- Validation was a concern in this case. I emphasized the need for tests that evaluate the robustness and generalization of the BDT in regions lacking training data, such as cross-checks using synthetic data, bootstrapping, or comparison with baseline cut-based selections.

Scope

- Provide a **short reference document** to allow **anyone** to be able to review analysis involving **"traditional" ML algorithm**.
- Provide a quick intro to these algorithm and to the associated technical terms

What it is

- A support document to help with analysis using "traditional" ML algorithm.
- A guide aimed for both reviewers and analyzers.
- Not a list of mandatory item.
- A draft, designed to evolve.

What it contains

Contents

1 Dataset bias 3

2 List of best practices 4

3 List of plots 5

4 Review checklist 6

5 Review feedback 7

A Glossary 9

B Introduction to Machine Learning 10
B.1 Quick Guide to Machine Learning . . . 10
B.1.1 General considerations . . . 10
B.1.2 Training and testing dataset . . . 10
B.1.3 Designing and training the algorithm . . . 11
B.1.4 Evaluating the performance of the algorithm . . . 12
B.1.5 Further optimization of the algorithm . . . 14
B.2 Algorithms Galore . . . 14

C Example analyses from CLAS12 18
C.1 Example 1 (good analysis): Lepton identification . . . 18
C.1.1 Description of the analysis . . . 18
C.1.2 Review checklist . . . 22
C.2 Example of analysis 2 (bad analysis): J/ψ event selection using dilepton invariant mass . . . 23
C.2.1 Description of the analysis . . . 23
C.2.2 Review checklist . . . 28

D Deploying AI/ML Tools 29

E Results of the initial survey on ML-based analysis reviews 31
E.1 Questions . . . 31
E.2 Answers . . . 31
E.3 Other Feedback: . . . 34

Best practices

- **Spurious correlations:** the algorithm relies on input features that are statistically correlated with the targeted class in the training data but are not causally related to the underlying physics.
- **Representation bias:** the range or diversity of the training samples is insufficient or the training data itself is unrepresentative of the data to which the algorithm will be applied.
- **Sampling bias:** the process of selecting training data systematically favors certain values or ranges of a physics observable, leading to an unbalanced or skewed dataset.

2 List of best practices

Certain recommendations may not apply to certain tasks and some recommendations are redundant. Use your best judgement when going against these recommendations and carefully explain why that is a sensible thing to do.

- Task Selection: Avoid tasks that have more potential to bias observables, due to the input features and training dataset having enough information to correlate classes to physics observables (e.g. for a multi-particle final state, prioritise single-particle PID over event selection).
- Physics-less Simulated Training Data: Avoid simulated training datasets created with given physics processes (e.g. for event selection, avoid multi-particle final states simulated with resonances, prioritise "flat" data).
- Uncontaminated Non-Simulated Training Data: Avoid non-simulated training datasets with large backgrounds or unfold the contributions from backgrounds [1] (e.g. do not create training sets from exclusive reactions with large backgrounds that are poorly understood).
- Feature Selection: Use physically motivated input features. If you are injecting physics in your training dataset, avoid input features that have enough information to extract your physics observables (e.g. for event selection with multi-particle final states simulated with resonances, avoid using input features where these resonances are identifiable).
- Feature Importance: Establish the input feature importance in a dataset [2], ensure the chosen input features are physically meaningful.
- Hyperparameter Selection: Report your algorithm architecture, its hyperparameters and justify the choice of algorithm and hyperparameters.
- Output Variable Correlation to Observable: Estimate the correlation between your algorithm's output and your physics observables.
- Introduce De-Correlation with Observables: The loss functions or training data can be modified to avoid correlations with physics observables [3]. Binning can also help remove the dependence on observables. De-Correlation techniques can make training slower and less likely to converge.
- Training Data Comparison: Ensure that your training data matches the data you apply the algorithm to, including variable ranges.
- Non-simulated Validation Sets: When training on simulated data, validate on both simulated and non-simulated data.
- Synthetic Validation Sets: Create test sets with different contributions from event sources (e.g. varying signal to background levels) to understand failure modes.
  If the input features in your training data are different to those in the data you apply your algorithm to (e.g. different ranges), create synthetic validation sets for various scenarios to understand the impact on the output of your algorithm.
- Input-Output Tests: Inject a known observable into your dataset, check that you are not distorting it by applying the ML-based algorithm. Check on a flat, physics-less dataset that you are not sculpting it.
- Impact on Fits: When fitting observables, check if your algorithm changes your fit parameters (e.g. peak fit width increases when increasing threshold on output).
- Comparison to Conventional Method: If possible, compare your results to ones obtained with a non ML method. Demonstrate where the differences come from.
- Vary Threshold on Output: In classification tasks, vary the threshold on your output to estimate the impact of the choice of threshold on your observables and the tests above.

Analysis checklist – for analyzers and reviewers

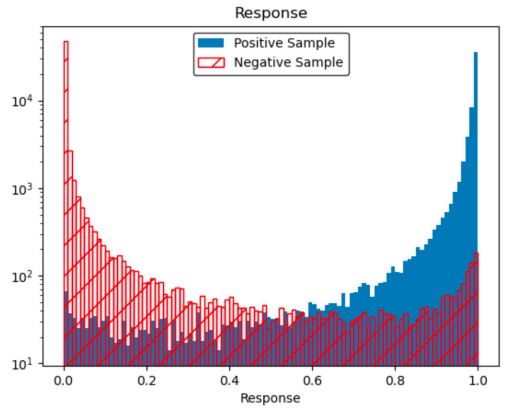
- Build in parallel to the best practice list
- General questions applicable to most ML algorithms
- Not an exhaustive list, to allow for discussion during the review process
- To be used during both the analysis and the review process

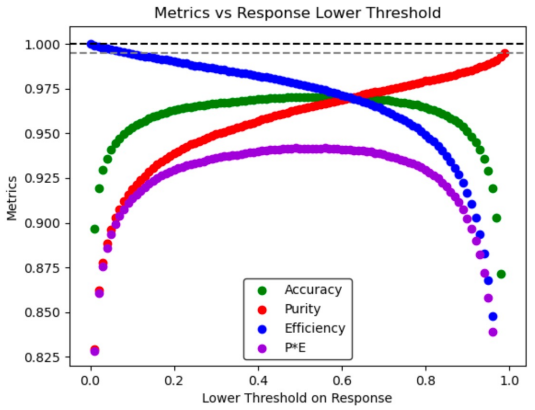
Review item	Questions	Validation
Task Selection	Does the task require AI/ML ? Can a simpler AI/ML algorithm be used instead ?	Y / N Y / N
Training Data	Is the training dataset representative of the data the algorithm will be applied to? Is the use (or not) of simulated data justified? If simulated, does it have physics in it? If not-simulated, does it avoid contamination?	Y / N Y / N Y / N Y / N
Feature Selection	Are the input features well motivated ? Was the feature importance evaluated?	Y / N Y / N
Algorithm and Metric definition	Can the algorithm and metrics be understood solely by reading the note ?	Y / N
Hyperparameter selection	Are the algorithm and its hyperparameters explicitly mentioned and justified ?	Y / N
Plots	Are the suggested plots present?	Y / N
Correlations to Observables	Are the correlations between algorithm output and observables evaluated? If needed, were de-correlation procedures applied?	Y / N Y / N
Validation Sets	Were metrics produced on non-simulated data? Was the algorithm applied to synthetic data? Were input-output tests done?	Y / N Y / N Y / N
Comparison to Conventional Methods	Is the comparison possible and present?	Y / N
Threshold Checks	Is the threshold on the algorithm output varied to evaluate the impact on kinematics, observables and fits?	Y / N

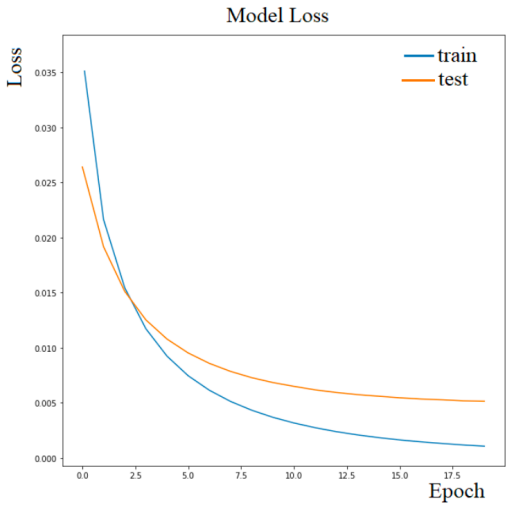
Analysis checklist – list of plots

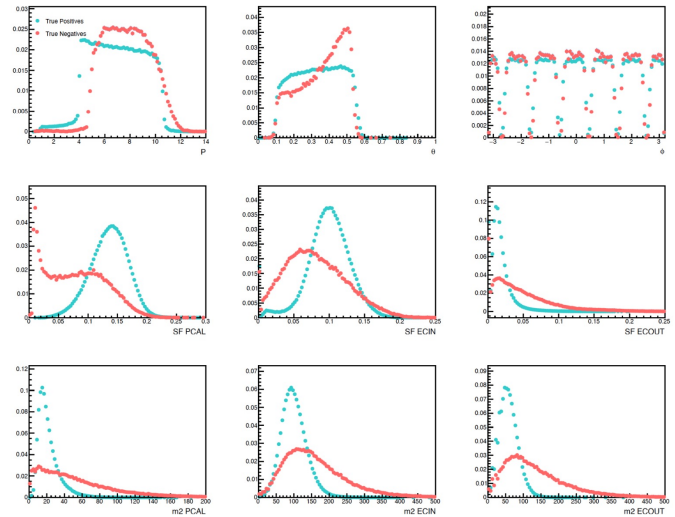
- **Inputs:**
  - 1D histograms of input features (by class for classification tasks).
  - Full correlation matrix between all inputs.
  - Separation power for each input variable (for classification tasks).
- **Algorithm Outputs:**
  - 1D histograms of output (by class for classification tasks).
  - 2D correlations between each input and output.
  - 2D correlations between output and relevant kinematic variables (e.g. P, θ, ϕ).
- **Training process:**
  - Loss vs epoch.
  - Overtraining check: train vs test output distributions (per class).
- **Dataset comparison:**
  - Input feature distributions for training/test data vs application data.
  - Algorithm output when applied on training/test data vs application data.
- **Performance metrics** (Classification):
  - Receiver Operator Curve (Signal efficiency vs Background rejection) curve.
  - Metrics vs output threshold.
  - Metrics vs relevant kinematics, at various output thresholds.
  - Input features, kinematic variables, and observables for accepted/rejected events, at various output thresholds.
- **Performance metrics** (Regression):
  - Prediction and truth vs relevant kinematics.
  - Metrics vs relevant kinematics.
  - Resolution (truth - prediction) vs relevant kinematics.

Some examples from the document



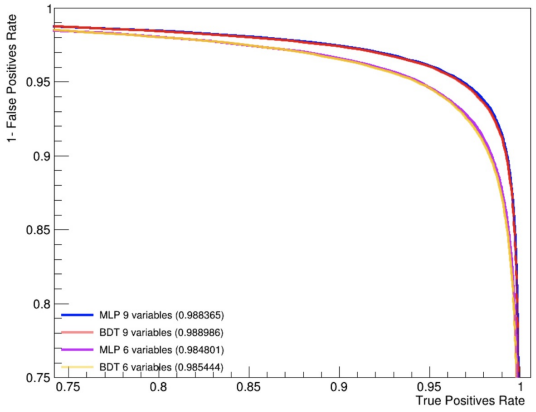


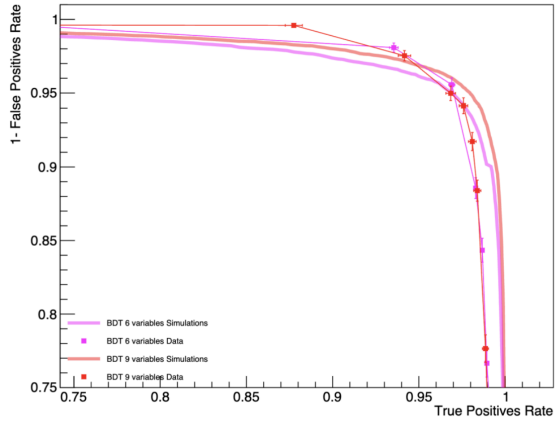




Examples

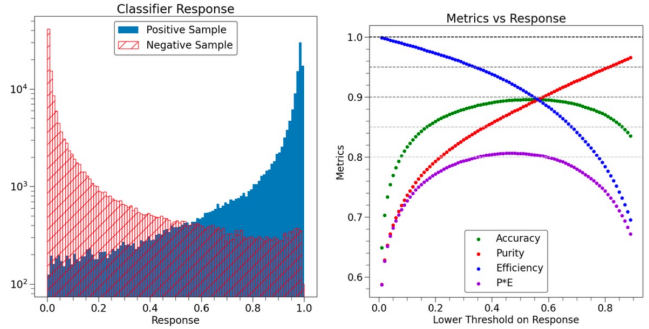
Example 1 – Lepton identification

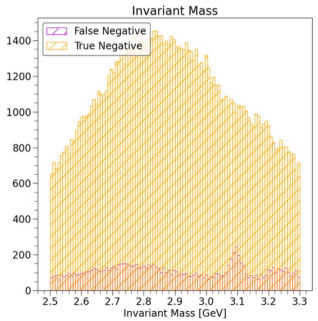


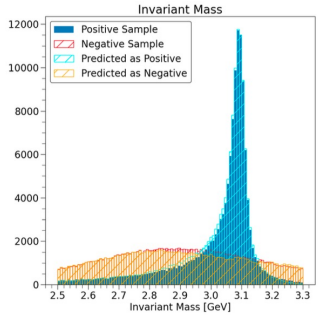


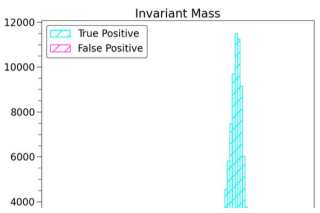
Review item	Questions	Validation
Task Selection	Does the task really require AI/ML ? Can a simpler AI/ML task be used instead ?	Yes No
Training Data	Is it representative of the data the algorithm will be applied to? Is the use (or not) of simulated data justified? If simulated, does it have physics in it? If not-simulated, does it avoid contamination?	Needs to use simulation. Training data is representative or corrected. Only single particles.
Feature Selection	Are the input features well motivated ? Was the feature importance evaluated?	Yes No
Algorithm and Metric Definition	Can the algorithm and metrics be understood solely by reading the note ?	Yes Yes
Hyperparameter Selection	Are the algorithm and its hyperparameters explicitly mentioned and justified ?	Mentioned, not justified
Plots	Are the suggested plots present?	Some
Correlations to Observables	Are the correlations between algorithm output and observables evaluated? If needed, were de-correlation procedures applied?	No but obviously not present. No
Validation Sets	Were metrics produced on non-simulated data? Was the algorithm applied to synthetic data? Were input output tests done?	Yes No No
Comparison to Conventional Method	Is the comparison possible and present?	Yes
Threshold Checks	Is the threshold on the algorithm output varied to evaluate the impact on kinematics, observables and fits?	Yes

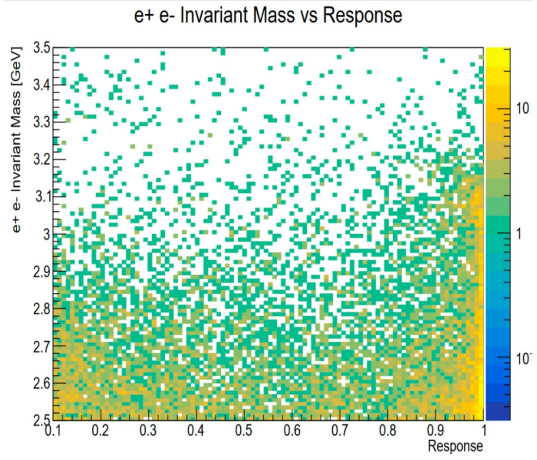
Example 2 – "Naïve" Jpsi analysis











Feedbacks and improvement

- This document is a draft…
  …and will evolve.
- Adapted from the short survey used to build this document.
- List of question to guide its improvement.
- To be filled by the review committee/analyzers at the end of the review process.

Question 1: Please describe the AI/ML tool or analysis that you were asked to review.

Question 2: Did you find this review particularly challenging compared to others without AI/ML components?

Question 3: Did you use these guidelines to review the analysis ? If so, did you find them useful ?

Question 4: Did you notice missing elements that should be included in the guidelines ?

Question 5: If relevant, please describe further questions and studies, beyond those in these guidelines, that you asked for when reviewing the AI/ML component of the analysis.

Question 6: Did you have trouble with understanding:

- The mechanics of the algorithm and technical terms?
- Figures or metrics presented by the analyser?
- The impact of the algorithm on the results?
- How to validate the algorithm?
- What to look for and what might be missing in the analysis?

Question 7: Any further comments?

Any questions or feedbacks ?