Standard Operating Procedure for integrating ML Classifiers into ANNA

Information needed from submitter

Version: February 4 2025

Last Updated by: Kelvin Kong

To integrate a new machine learning classifier into ANNA, we need the following resources from the classifier development team.

1. Original algorithm/model and its source codes
2. List of required variables and test(s)
3. Logics for running model(s)
4. Preprocessing rules
5. Validation data table(s)
6. Contact Person(s)
7. Manuscript and Description Text(s)
8. **Original algorithm/model and its source codes**

The items listed in this section should contain all the required items to run the classifier.

* The model(s) itself.
* Sample script(s) to load, initialize, run the model, result visualization and presentation.
  + For Python native models, specify arguments needed (If applicable)
  + For non-Python native models, specify the method to call/initialize and run the model within Python.
* A list of required software, packages and modules, along with their version numbers.
* Python Virtual Environments (If applicable)
* Expected outcome(s) of the model, and how it converts to human readable results.
  + Examples:
  + For binary classifiers, machine generated result could be:
    - 0, and 1.
    - Human readable result could then be: 0 represent Negative/Not significant, and 1 represent Positive/Significant
  + For multi-level classifiers, it could be:
    - 0 represents outcome 0, 1 represents outcome 1, 2 represents outcome 2 etc.
  + For classifiers with probabilities of outcome, include how to extract the probabilities.
* GitHub Repository Link(s)
* Any other items required to run the model(s) locally (If applicable)

1. **List of required variables and test(s)**

For machine learning classifiers that uses blood work test results should provide:

* All the required analytes for running predictions.
* The analytes’ exact name required in the model. (HCO3 vs Bicarbonate etc.)
* Analytes’ data type (Numerical, categorical variables, free text etc.)
* The units of the analytes.
* The order of the analytes in the dataset. (If applicable)
* The rules of forming a valid data row for prediction.
  + We support combining different test panels from different test dates to become a valid row of data. However, it must specify the range of dates allowed to be considered.
  + For example: It can be:

1. CBC (A fixed date) + Chemistry Panel (+/- 5 days)
2. CBC (+/- 2 days) + Chemistry Panel (+/- 5 days) + Urinalysis (+/- 5 days)
3. CBC (A fixed date) + Chemistry Panel (A fixed date) + Urinalysis (A fixed date) + Immunology (+/- 3 days)

* Maximum number/ratio of analytes can be empty/missing for running the classifier (If applicable)
* Analytes that must be present for running the classifier (If applicable)

1. **Logics for running classifier(s)**

For all classifiers, we require a set of logic which determine when and how the model should be run, including:

* If the classifiers require a prior test ordered/completed to be eligible for prediction.
  + For example, Leptospirosis Classifier requires either 8245 Leptospira Panel, Canine (CAHFS) Test/8780 PCR - Leptospira Sp./8615 LEPTO (Antibody Lateral Flow) within 7 days of CBC test date.
* If the classifiers require further information which a normal database doesn't provide.
* The logic of running a classifier if the patient data is missing some of the variable(s).
* For classifiers with more than one algorithm, we require a rule to show the sequences of running the algorithms (If applicable), or the conditions to run individual algorithms.
  + For example, if a classifier consists of two algorithms, a generic algorithm, and a more refined algorithm, to detect the same disease, we would like to know what the conditions are to run the generic algorithm, and what are the triggers to run the refined algorithm.

1. **Preprocessing rules**

It is helpful to know the rules to preprocess the eligible patient data and being fit into the machine learning classifier. The following items are required if applicable.

* Scaling of numeric values, including any value transformation, normalization etc.
* Rounding/truncating of numeric values with decimals.
* Encoding of categorical values. (OneHotEncoding etc)
* Standardizing of numeric, or categorical values.
* Removal of abnormal values/outliers.
* Rules to handle missing values in predicting data sets. (Fill NAs with 0/Impute using Mean/K Nearest Neighbour etc)
* Other data preprocessing methods required before fitting into the classifier.

1. **Validation data table(s)**

The purpose of the validation data table is to check if the model behaves as expected when being integrated to Anna. It also serves as the sample data set being used to test run before moving into deployment.

***Terminology definition:***

Types of conditions/data:

1) Disease status of patient (affected/not affected) – real world/ultimate ground truth

2) Classifier prediction – stand alone, as assessed by developer

3) Classifier prediction – by Anna

In this context:

Validation refers to assessing the concordance between the results generated by ANNA (3) and those determined by the classifier's developers (2). Validation of the classifier itself was performed by its developers and is not within the scope of this validation.

Ground truth refers to the classifier prediction supplied by the developer (2) and NOT the gold standard/disease status of the patient (1).

Hence, we are looking for the validation data table(s) that contains:

* The values of the analytes being used for model validations, and the corresponding expected model output(s).
  + We are interested in seeing if the model runs as expected after integration. Hence, an expected prediction result(s) from the model algorithm(s) is required rather than having ‘golden’ ground truth (Type 1).
* Preferred file format: CSV/XLSX/JSON/XML Tabular style

If the model is trained using patient data from UC Davis VMACS, please include the following meta data and identifiers:

* PatientID
* Test Date(s)
* List all test dates if multiple tests are required to form a data row.
  + E.g.: Complete Blood Count Date: (2024-10-31), Small Animal Panel 2 (2024-10-30)
* VisitID(s) (Optional)
* TestID(s)
  + List all test IDs that were used to combine for validation

1. **Contact Person(s)**

It would be grateful if we have a contact person that we can further inquire details of things we are not sure. Name and Email is sufficient.

1. **Manuscript and Description Text(s)**

We would like the following items in this section:

* Published Manuscript link/DO
* A short summary description of the classifiers, including
  + Background
  + Methods
  + Results