**Abstract**

In recent years there has been significant development on interpretability methods [1]. Some methods try to tackle the idea of "why?" instead of "what?" from an axiomatic approach [2] while some do so from a visualization perspective of real-world data [3]. There seems to be a knowledge gap on how to effectively evaluate and validate any given interpretability method.

We offer a framework for evaluating an interpretability method. Our approach can be summarized in the following scheme [Figure1]:

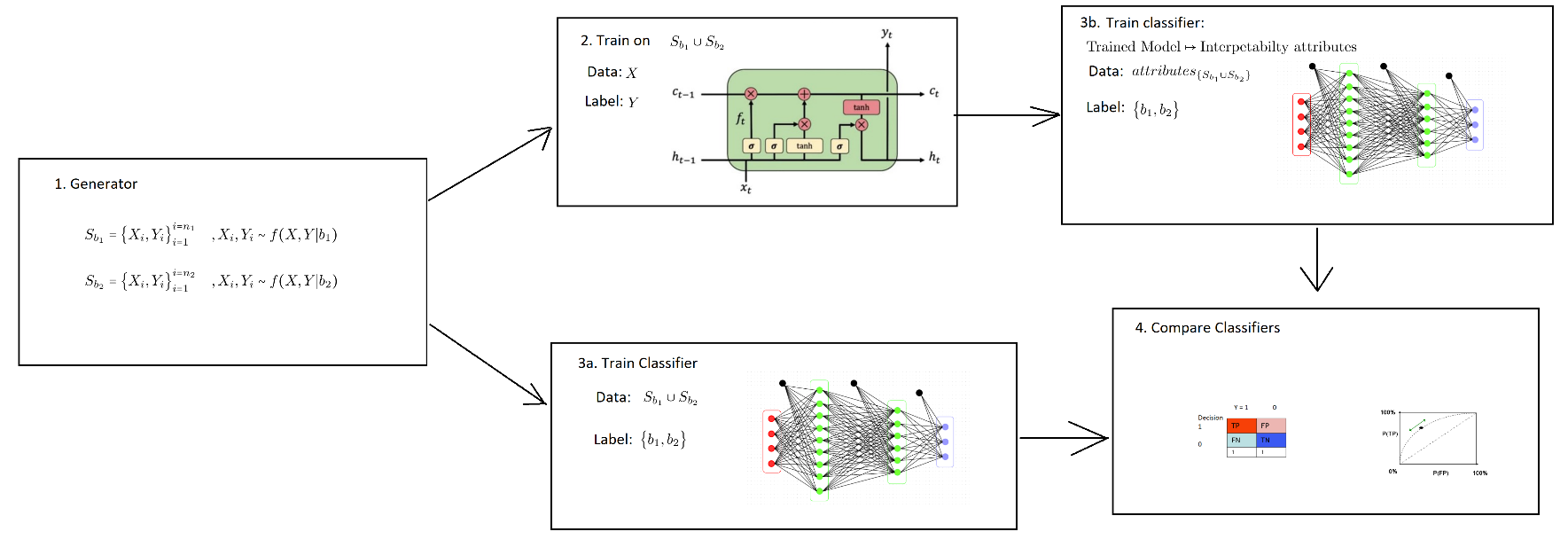
1. Generate a 2-modal data-set.
2. Train a machine learning model on the data set.
3. Train two classifiers for the modalities of the data set:
   1. A *clean* view – training only with the data set.
   2. An *interpreted* view – training only with attributes of the trained model.
4. **Evaluate interpretability method by comparison of the *clean* classifier to *interpreted* classifier with simple classification metrics.

Figure Evauation Scheme

We used this framework to evaluate the *Integrated Gradients* [2] interpretability method. Our data sets were taken from the hydrology domain, we basically took two synthetic (but realistic) rainfall and potential evaporation (PET) data series, representing two weather scenarios (step 1 data). They were merged and fed a rainfall-runoff model that simulated streamflow (step 1 labels). We trained LSTM model (step 2) and run several interpretability methods and then trained the *clean* and *interpreted* classifiers (steps 3a+3b). Finally, we used classical ML classification metrics like: *accuracy*, *precision*, *recall* etc. to evaluate which method is better comparatively.

The reasoning for using real scientific data models is also to validate the summary of the *Integrated Gradients* method. In a sense the *story* that the attributes tell can be revised and corroborates through our hydrologic expertise. [TBD Efrat]

**References:**

1. Christoph Molnar. Interpretable Machine Learning. A Guide for MakingBlack Box Models Explainable. https://christophm.github.io/interpretableml-book/. 2019.
2. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution  
   for Deep Networks". In: CoRR abs/1703.01365 (2017). arXiv: 1703.01365.  
   url: <http://arxiv.org/abs/1703.01365>.
3. Amy McGovern et al. "Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning". In: *Bulletin of the American Meteorological Society* 100 (Aug. 2019). doi: 10.1175/BAMSD-18-0195.1.