2nd European Photon & Neutron EOSC Symposium

26 October 2021

Machine Learning-based Spectra Classification

Yue Sun; Sandor Brockhauser; Christian Plueckthun; Zuzana Konopkova

European XFEL; University of Szeged

26 October 2021

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.

Background

❑ Background (Big data):

- Experiments in photon sciences at synchrotrons or XFELs (e.g. SCS, FXE and HED experiments) always generate a large volume of data.

❑ ML algorithms and its requirements:

- A great amount of data which are clearly annotated and complete for covering the requirements of scientific reuse.

❑ ML, Big Data requirements on Ontologies, and Application Definition:

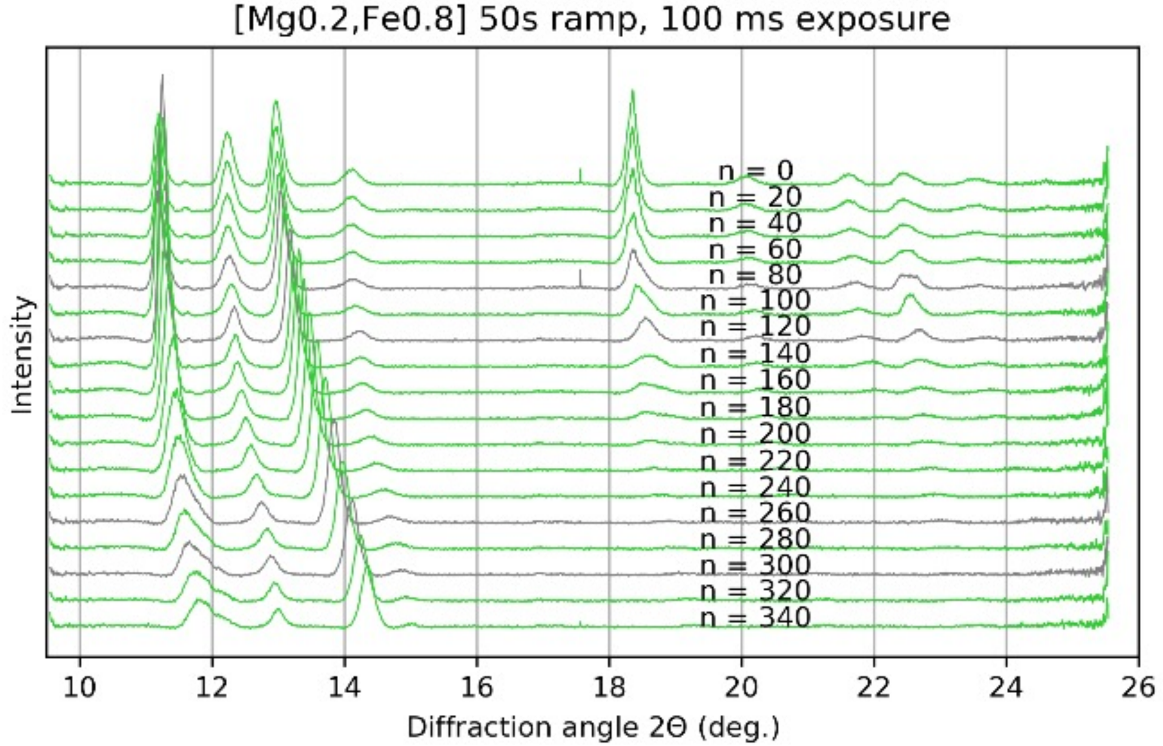
- The research outputs should align with the '**FAIR**' principles, meaning that data, software, models, and other outputs should be *Findable, Accessible, Interoperable,* and *Reusable*;
- For data re-use, the metadata shall follow a scientific experiment data model; Appropriate NeXus application definitions shall be defined.

panosc
photon and neutron open science cloud

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services

Example of Data Reuse

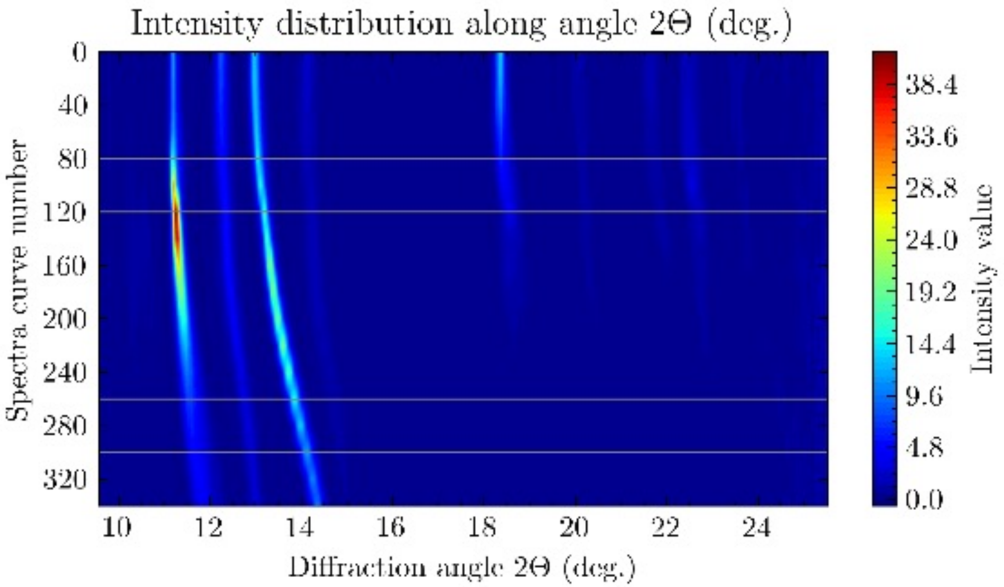
HED Diffraction Spectra data



Data set: 349 samples with each of 4023 features

Two phases: Low and High pressure

Training data: 4 (original)+40 (simulated) = 44

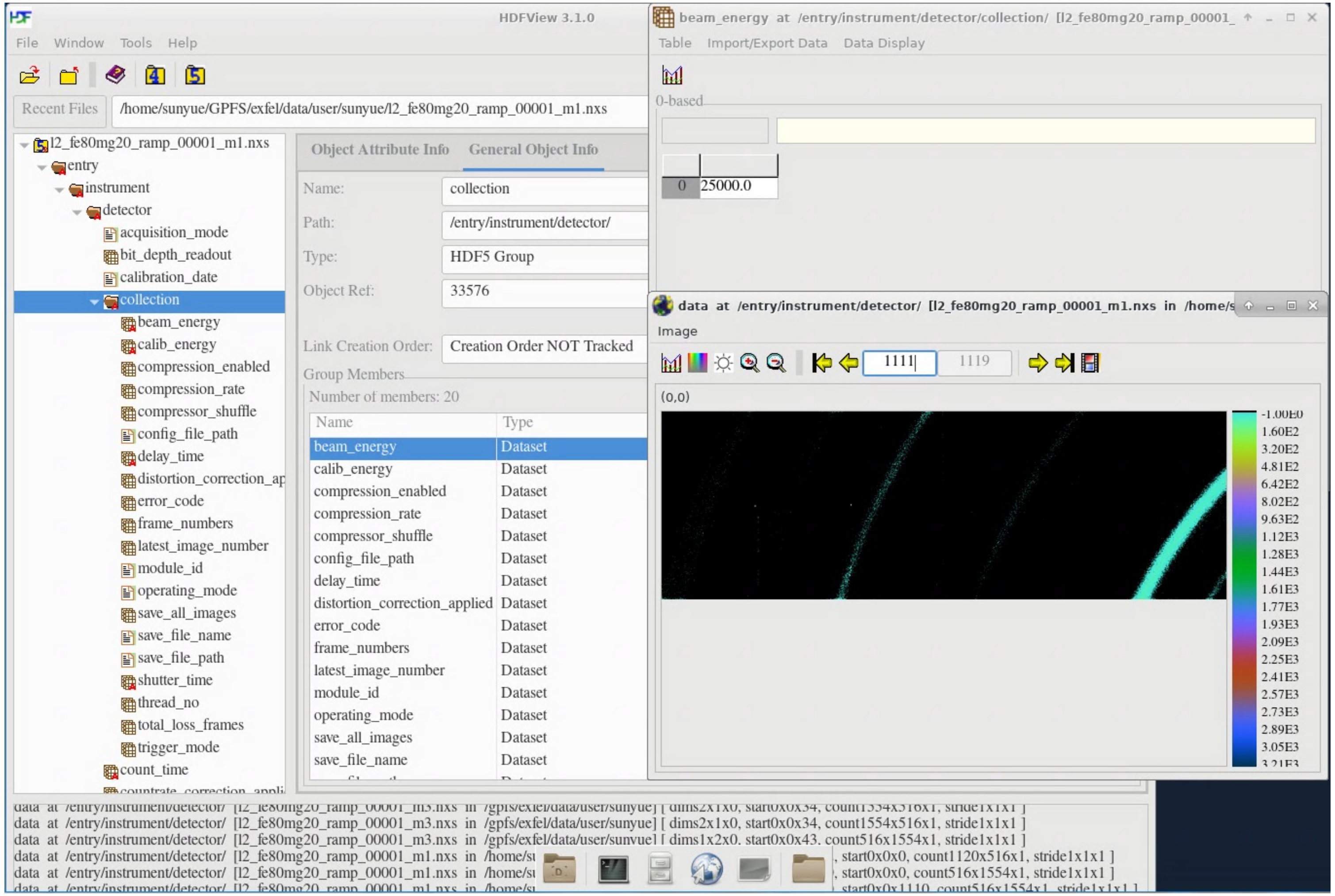


Measured at PETRA III beamline P02.2 using a 25.6 keV beam.

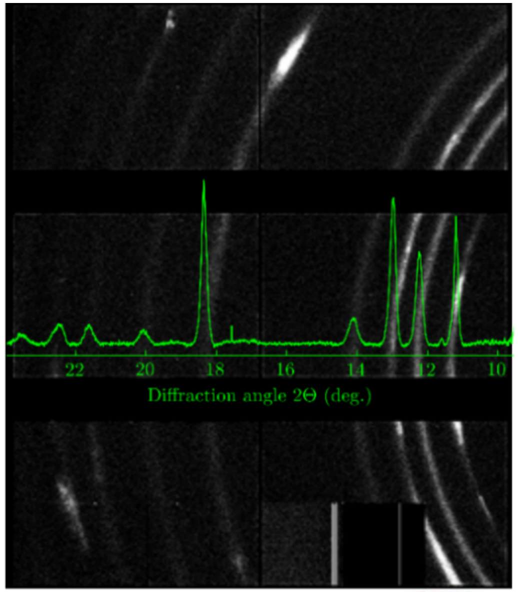
The **major spectral changes** we aim to capture are

- the change of intensity distribution (e.g. drop or appearance) of peaks at certain locations, or
- the shift of those in the spectrum.

NeXus Raw data file example



- NeXus base classes are used, but no application definitions (only free key-values under **NXcollection**).
- No real experimental data was registered (e.g. **beam_energy**).

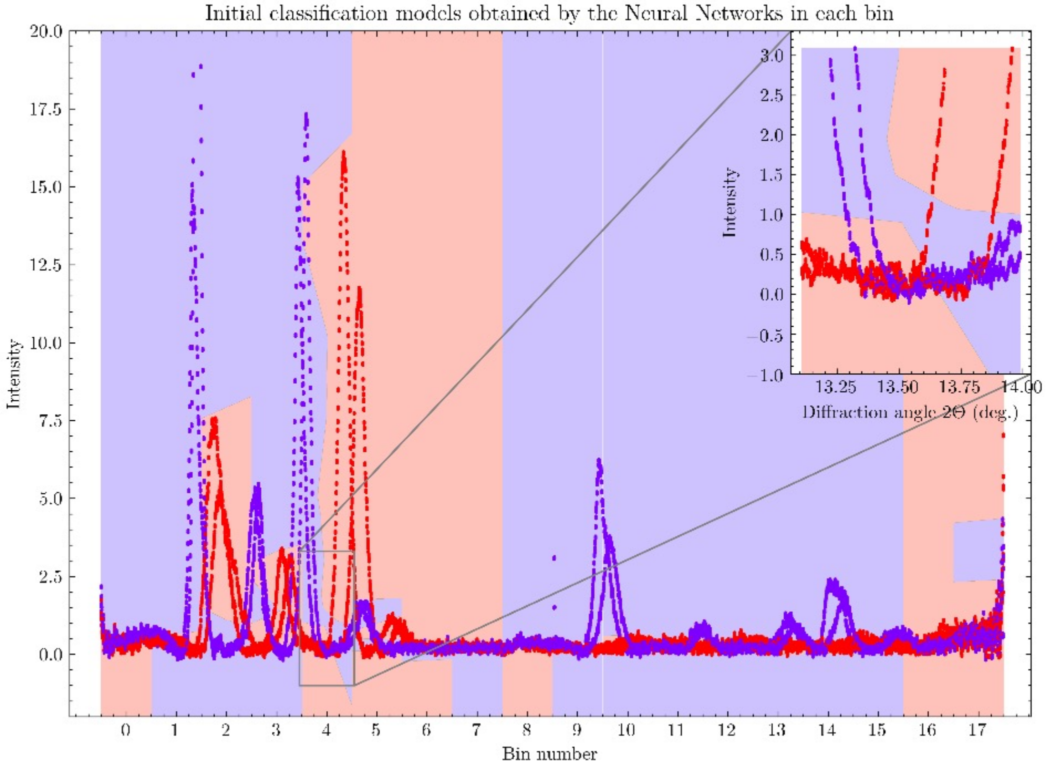


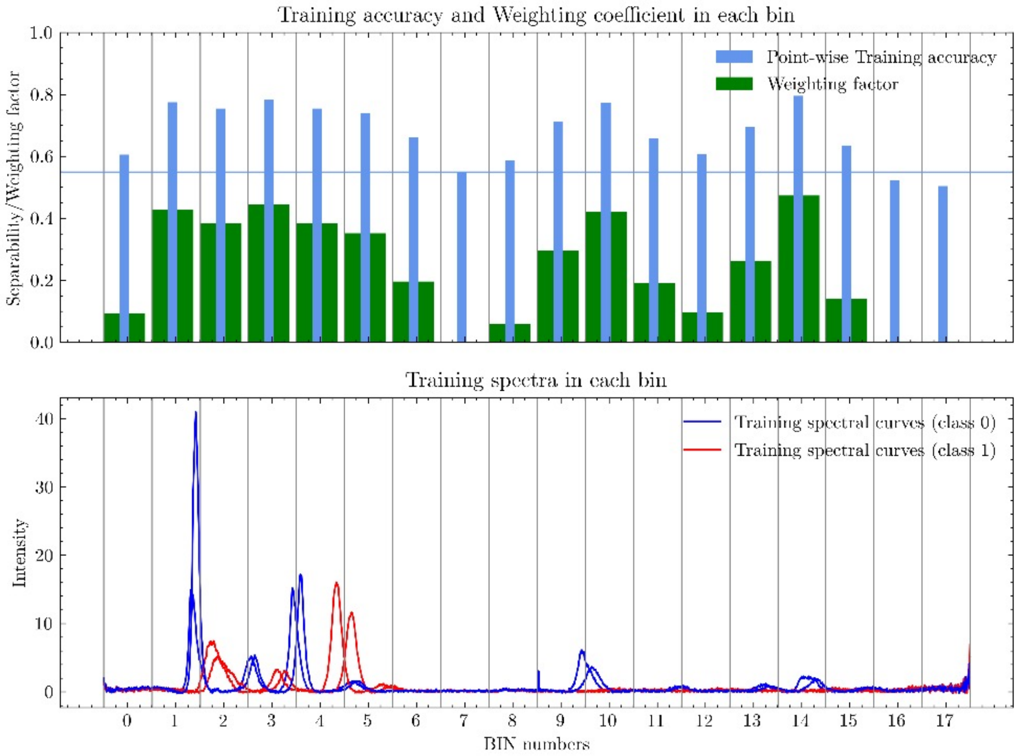
2D diffraction pattern

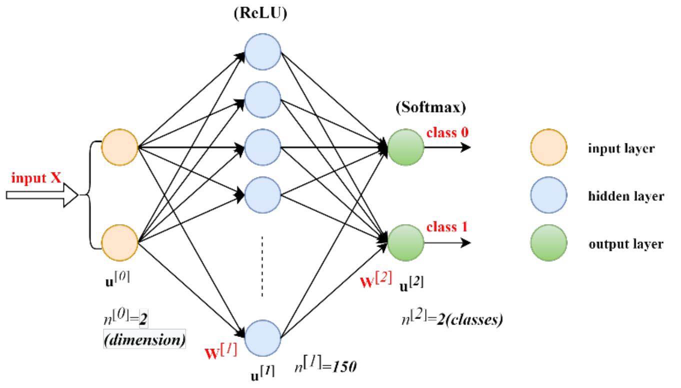
One of the NeXus Raw data file displayed in HDFview software

Two-layer NN with Weighting technique for spectra classification

In this work, the spectra classification task is regarded as a general two-dimensional (2D) space segmentation problem.



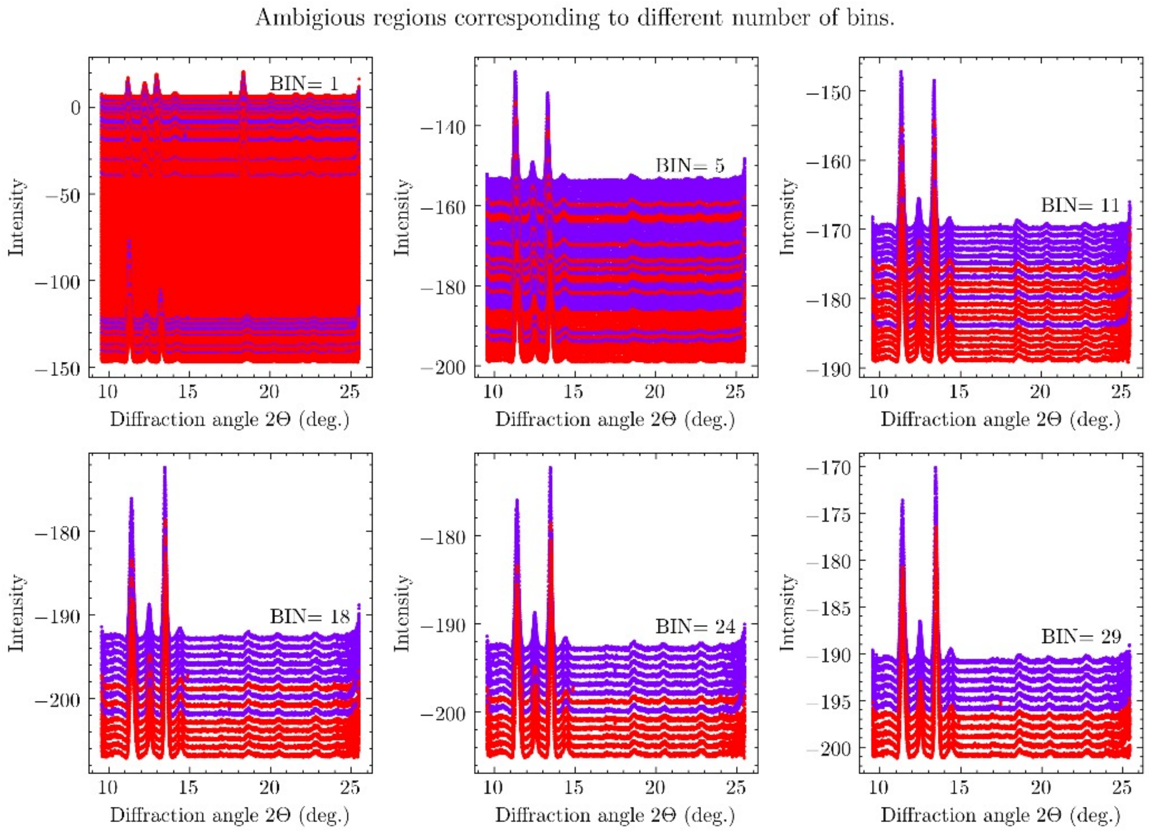


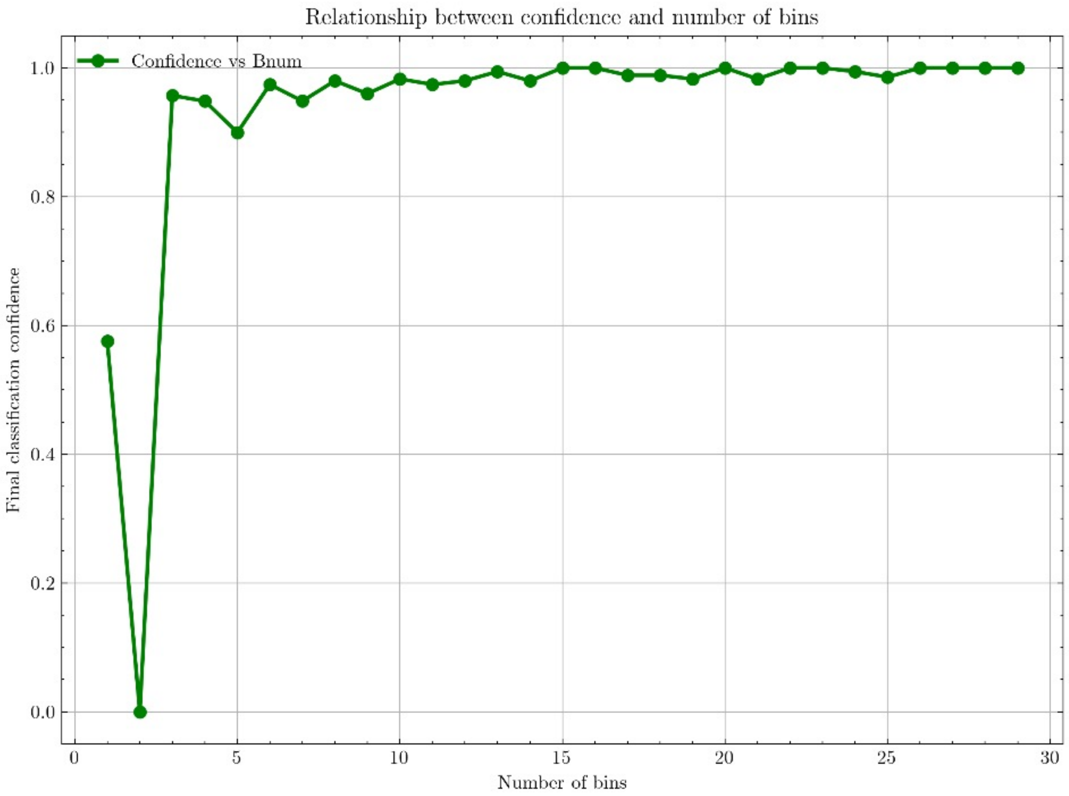


The distribution of point-wise training accuracy (separability), and the believability weighting factor for each bin.

Believability weighting factor:

$$w_b = \max\left(0, 1.4337{A_b}^2 - 0.4337\right)$$





$$L_{curve} = \underset{k \in \{0,\cdots,C-1\}}{\operatorname{argmax}} \sum_{b=1}^{B} w_b \sum_{i=1}^{N_B} \hat{C}_k^{(i)}$$

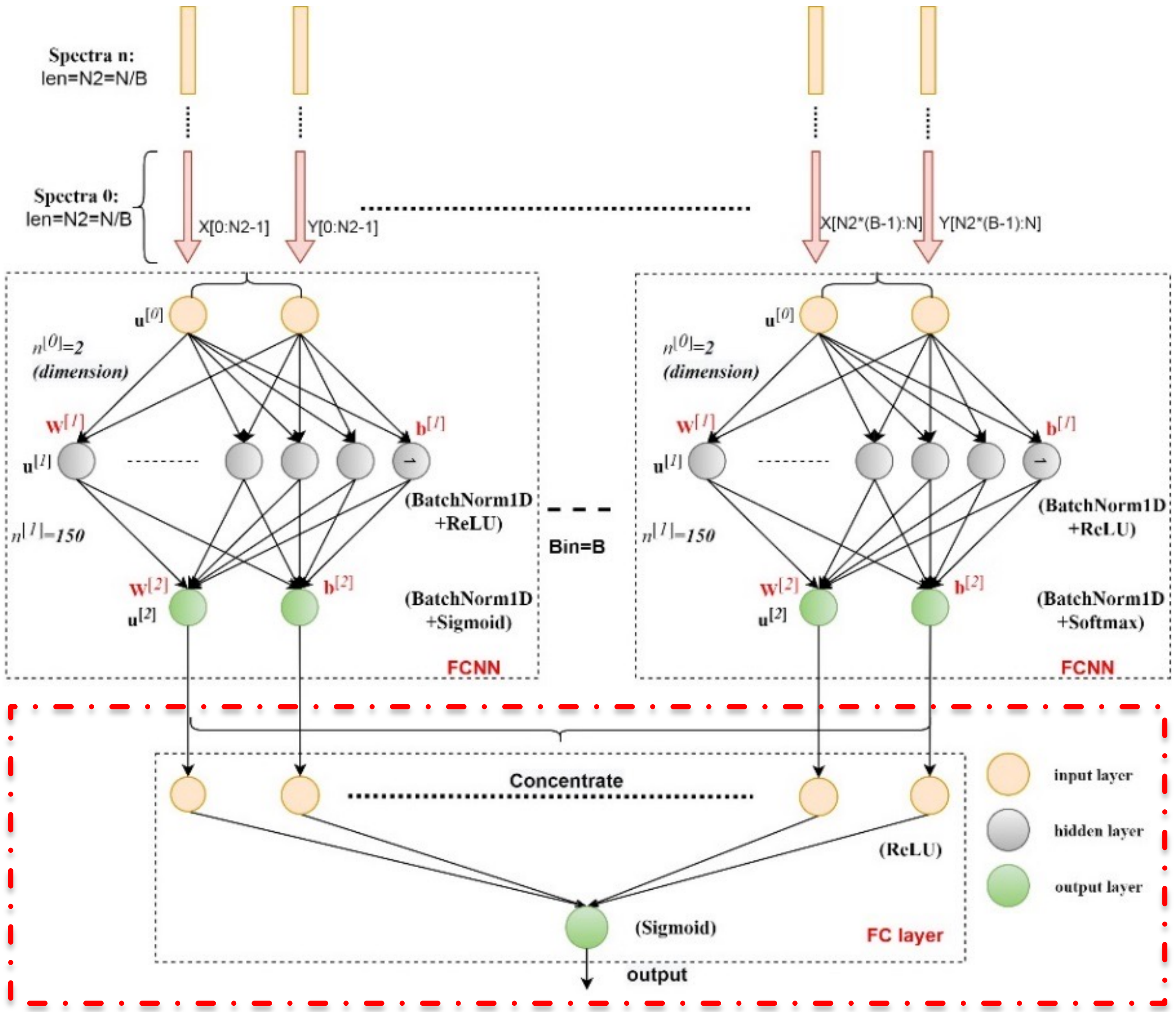
Performance metrics: $P_{conf} = 1 - \frac{N_f}{N_t}$

- Submitted a paper ‘Machine Learning Applied for Spectra Classification in XFEL Sciences’ to Data Science Journal;
- https://github.com/European-XFEL-examples/panosc-ml-spectra-classification.

panosc photon and neutron open science cloud

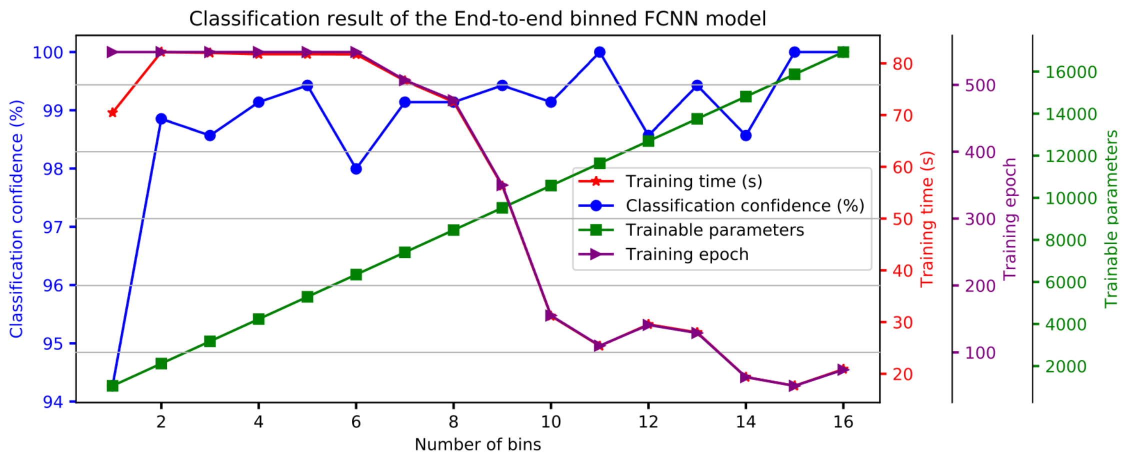
ExPaNDS European Open Science Cloud Photon and Neutron Data Services

End-to-End FCNN with automatically capturing weighting factors model



End-to-end FCNN with automatically capturing weighting factors model

Classification results under different number of bins:



As the number of bins increases:

Classification confidence increases;

Trainable parameters increase linearly;

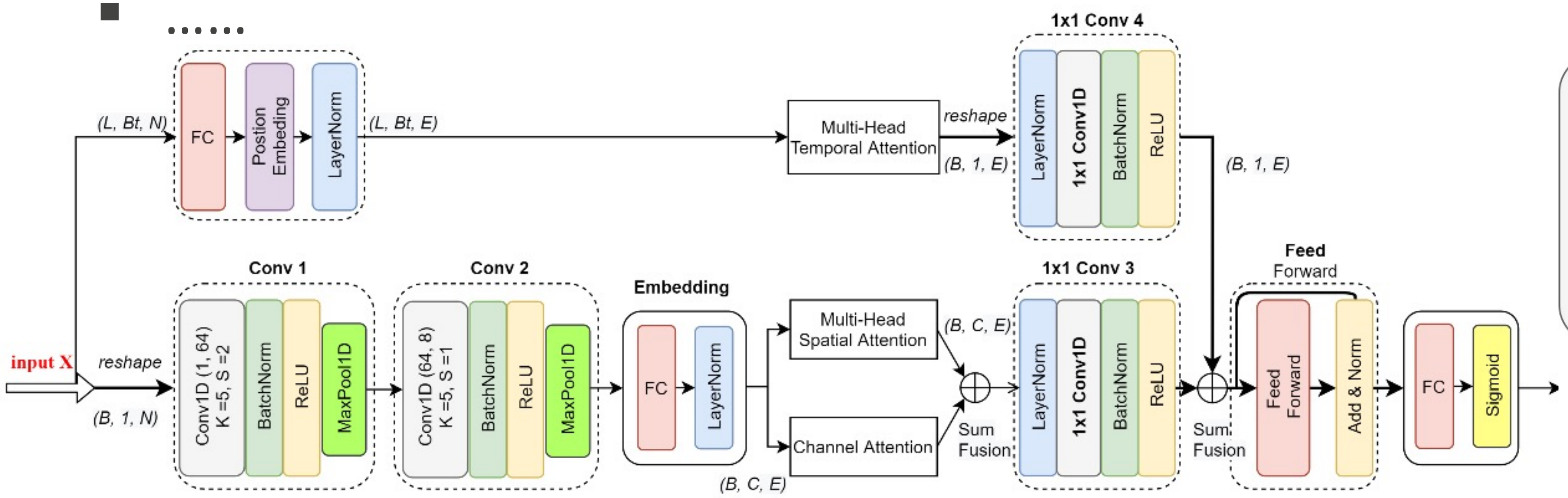
Training epochs decreases;

Training time decreases.

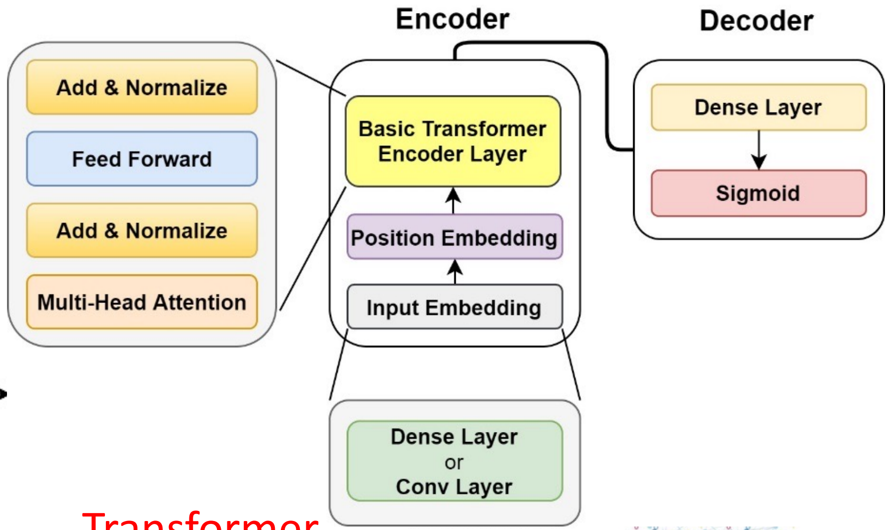
Other classification models

Under the setting of 1D spectral time series classification:

- 1D Fully Connected NN (FCNN)
- Convolutional Neural Network (CNN) solution, ResNets solution
- LSTM-based solution
- Transformer-based solution
- CNN+Multi-head Attention
- ……



CNN+Attention



Transformer

Conclusion and Perspectives

- The process of selecting/creating the training set is still limited, because the data is not properly annotated and some key information (such as the pressure value for each diffractogram) is missing.
- Appropriate NeXus Application Definitions needs to be developed.
- PaN portal is enabling 'FAIR' principle:
  - Data search;
  - Download / Data Analysis and Visualization cloud environment;
  - Automatic data interpretation via NeXus ontologies for
    - interoperability, and
    - reusability

Thank you

Yue Sun (European XFEL, University of Szeged):
yue.sun@xfel.eu

PaNOSC and ExPaNDS projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 823852 and 857641, respectively.