A First Application of Collaborative Learning in Particle Physics

29th November, 2021

Stefano Vergani HEP GROUP – Cavendish Laboratory

OVERVIEW

- Introduction
- Collaborative Learning and the Colearn Library
- Usability in Particle Physics
- Neutrino Dataset
- Keras Models
- Results
- Conclusion and Future work

Personal Introduction

- PhD student specialised in neutrino physics and artificial intelligence
- Funded by UK STFC Centres for Doctoral Training in Data Intensive Science
- Working on data analysis for Deep Underground Neutrino Experiment (DUNE) and its prototype Proto-DUNE Single Phase (SP)
- Focus on new technologies for AI: hardware and software
- Currently finishing internship at UK Cambridge-based company Fetch.ai (https://fetch.ai/)
- Co-author Attila Bagoly works at Fetch.AI
- For any questions: sv408@hep.phy.cam.ac.uk

Federated Learning and Collaborative Learning

Federated Learning: the model is sent to the data (typically edge devices) and trained locally.

Example: Apple trains their model to recognise the user's writing style directly on the I-Phone, without actually knowing what the user writes.

Collaborative Learning: there is a collection of different datasets. One dataset is randomly picked to see if the weights from the chosen dataset improve the performance of the model.

Example: the user wants to sell on the market validated images to improve one machine learning model. The potential buyers want to see how good are these images compared to others.

Colearn Library

Colearn is a library that enables privacy-preserving decentralized machine learning tasks on the FET network.

Useful for building a shared machine learning model without:

- relying on a central authority
- Revealing your dataset to the other stakeholders

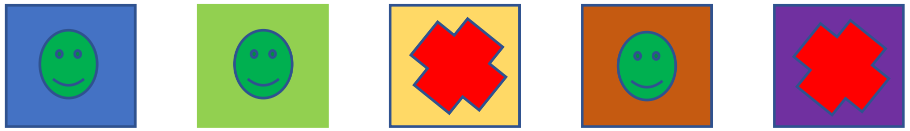
The library can be found here: https://docs.fetch.ai/colearn/

Colearn Library in a Nutshell

A group of learners come together and each one of them has a dataset

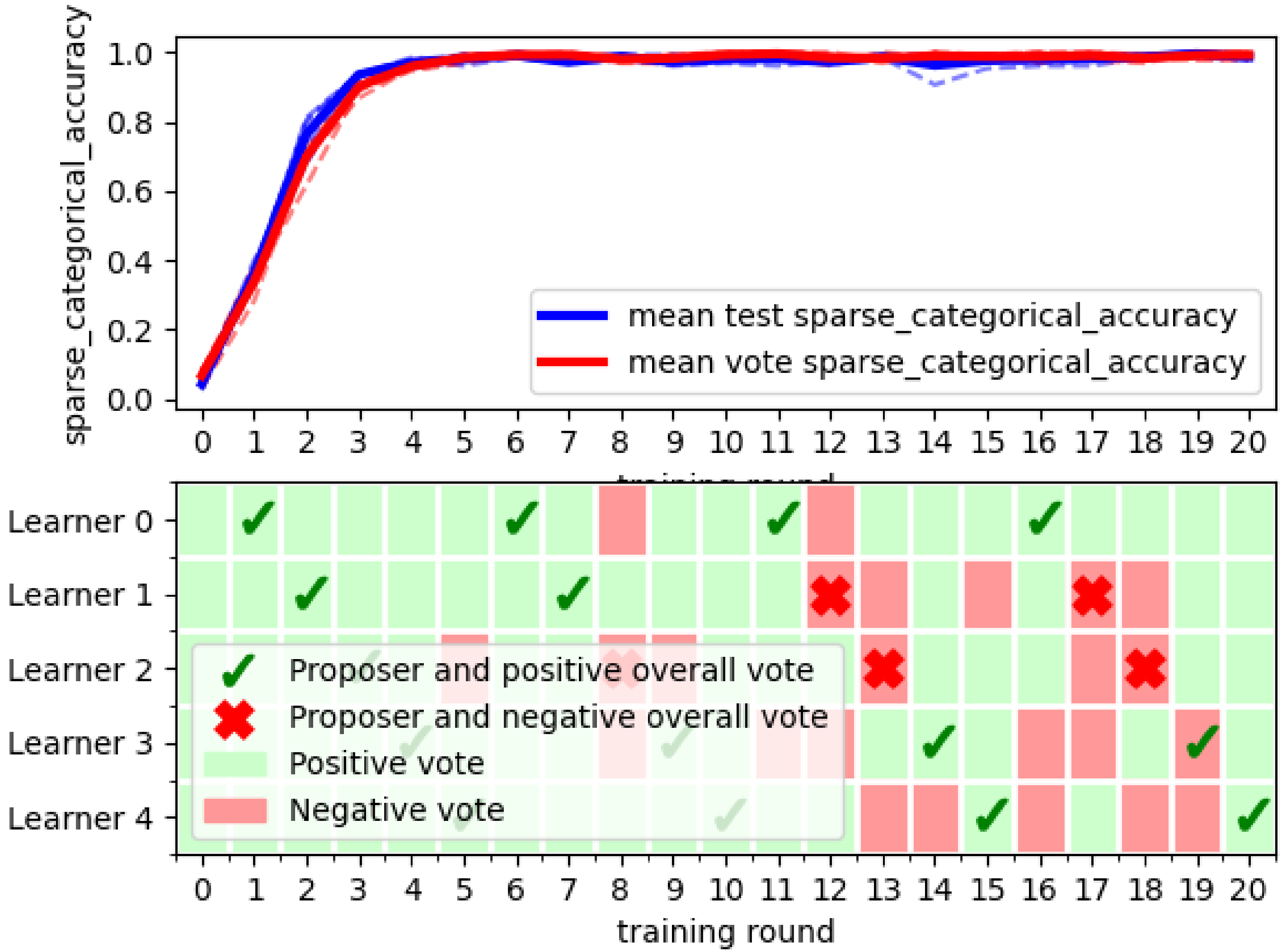
One learner is randomly selected to train the model

Each learner votes on whether the new weights are an improvement



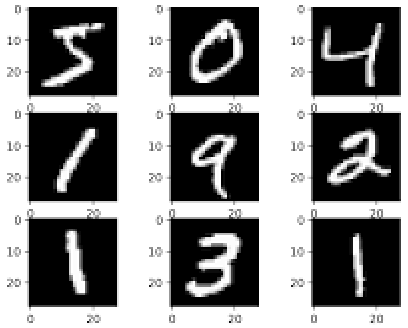
If the majority agrees, new weights are accepted, and a new round starts

Colearn for Keras with MNIST dataset



5 Learners

20 Rounds



MNIST Dataset

Usability in Particle Physics

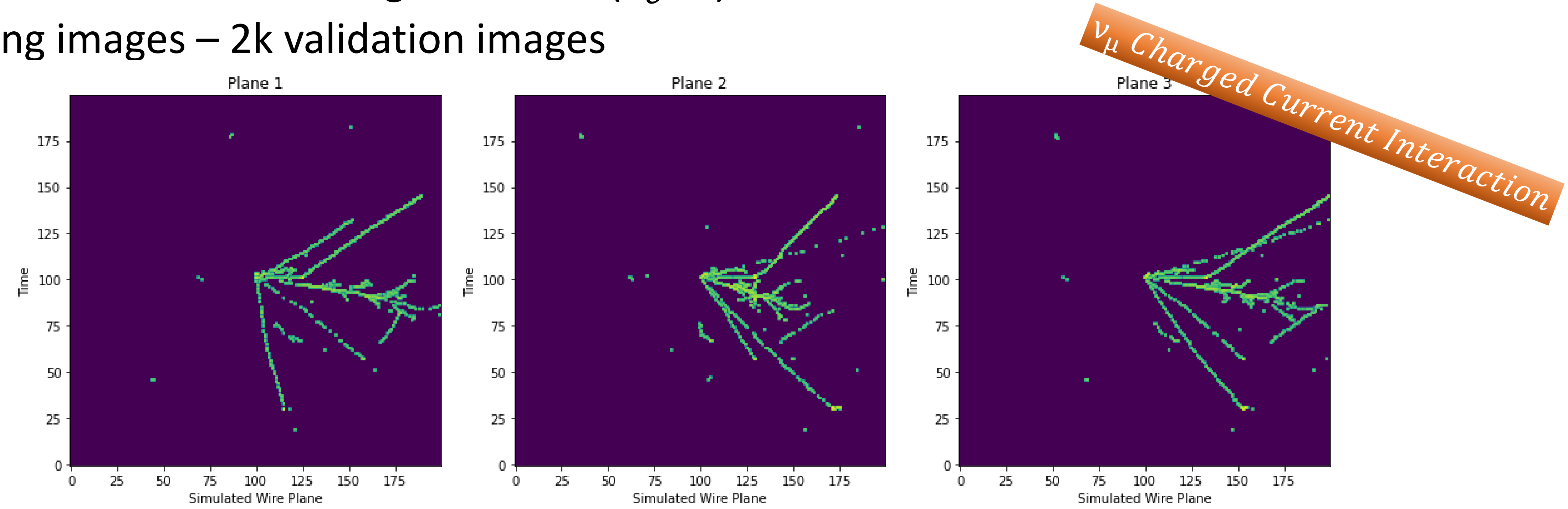
Only for DUNE, there are countless different datasets and machine learning (ML) models being used.

This library can be used for:

1. Using different datasets to train together a ML model.
2. Identifying the most performant datasets for a given ML model scientifically and impartially.
3. Identifying the most performant ML model for a given dataset. It is possible to change the structure of the code so that each learner will bring a different ML model instead of a different dataset.

Neutrino Dataset

- Liquid Argon Time-Projection Chamber (LArTPC) simulated images
- 3 simulated wire planes, 200x200 pixels which mimic wire readouts
- 3 classes of interaction: neutrino neutral current (NN) , muon neutrino charged current (ν_μCC), electron neutrino charged current (ν_eCC)
- 10k training images – 2k validation images

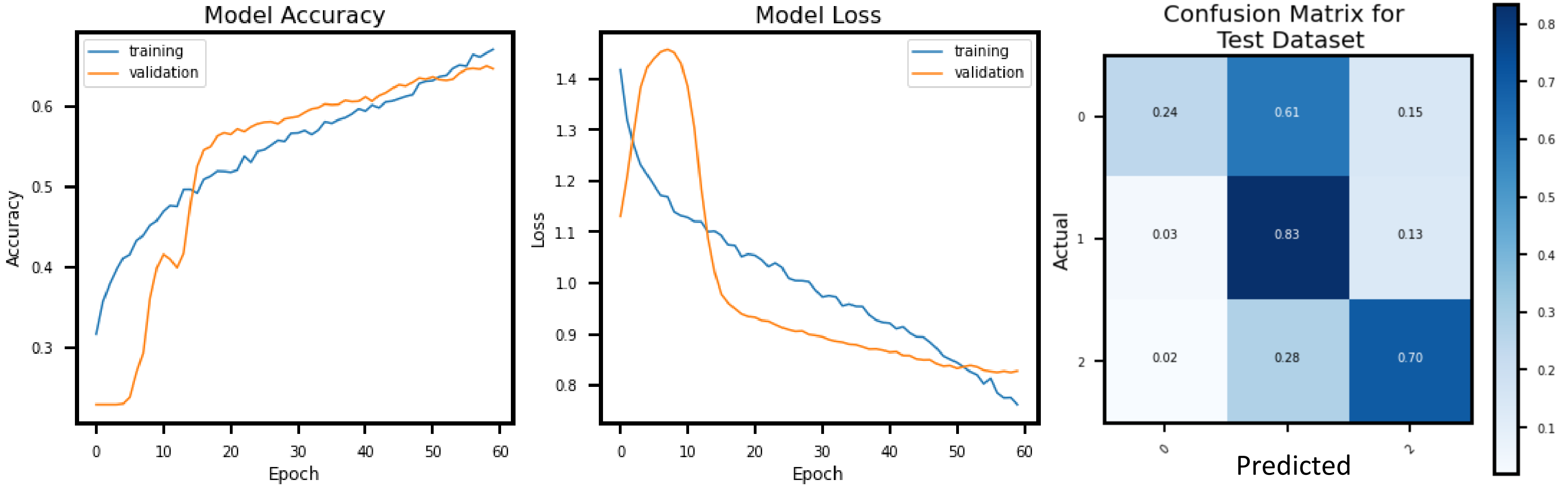


Neutrino Dataset for Learners

- For this experiment, we broke the neutrino train dataset into x parts and each part was assigned to one learner. This is to simulate different datasets coming from different sources.

- To maintain the proof of concept as objective as possible, the test dataset has been divided into x parts as well. Each learner will test the new set of weights on its own subset of test images, as it would normally happen.

ResNet 50 V2 and DenseNet 169

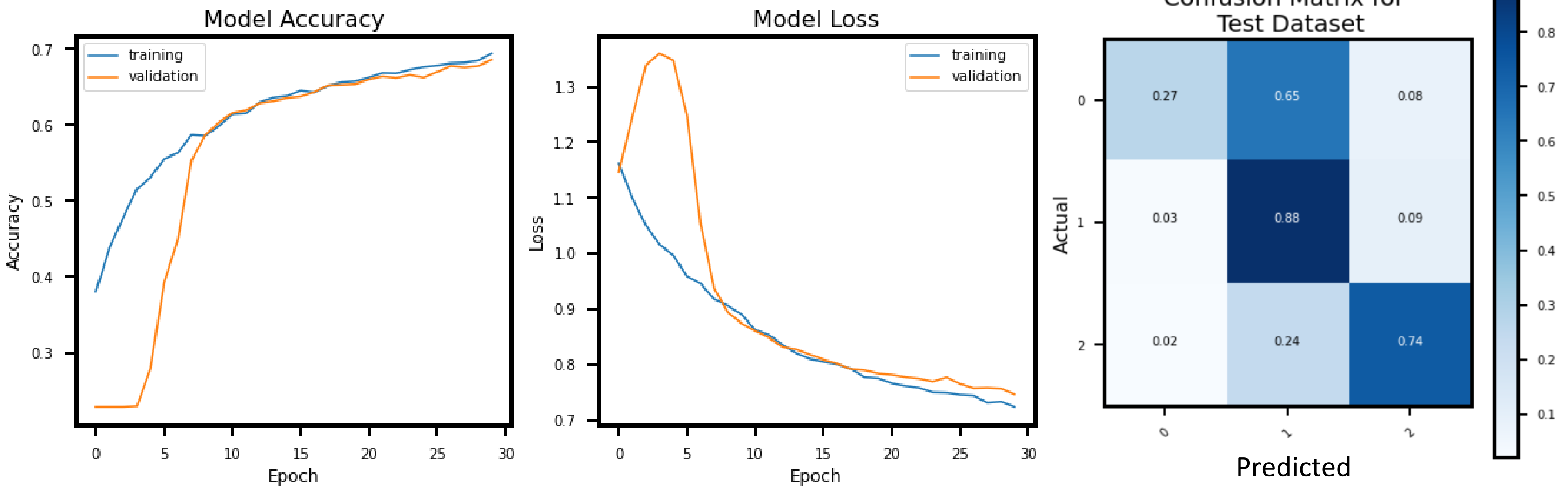


ResNet-50 V2

Validation accuracy **64.15%**

Confusion Matrix Values:

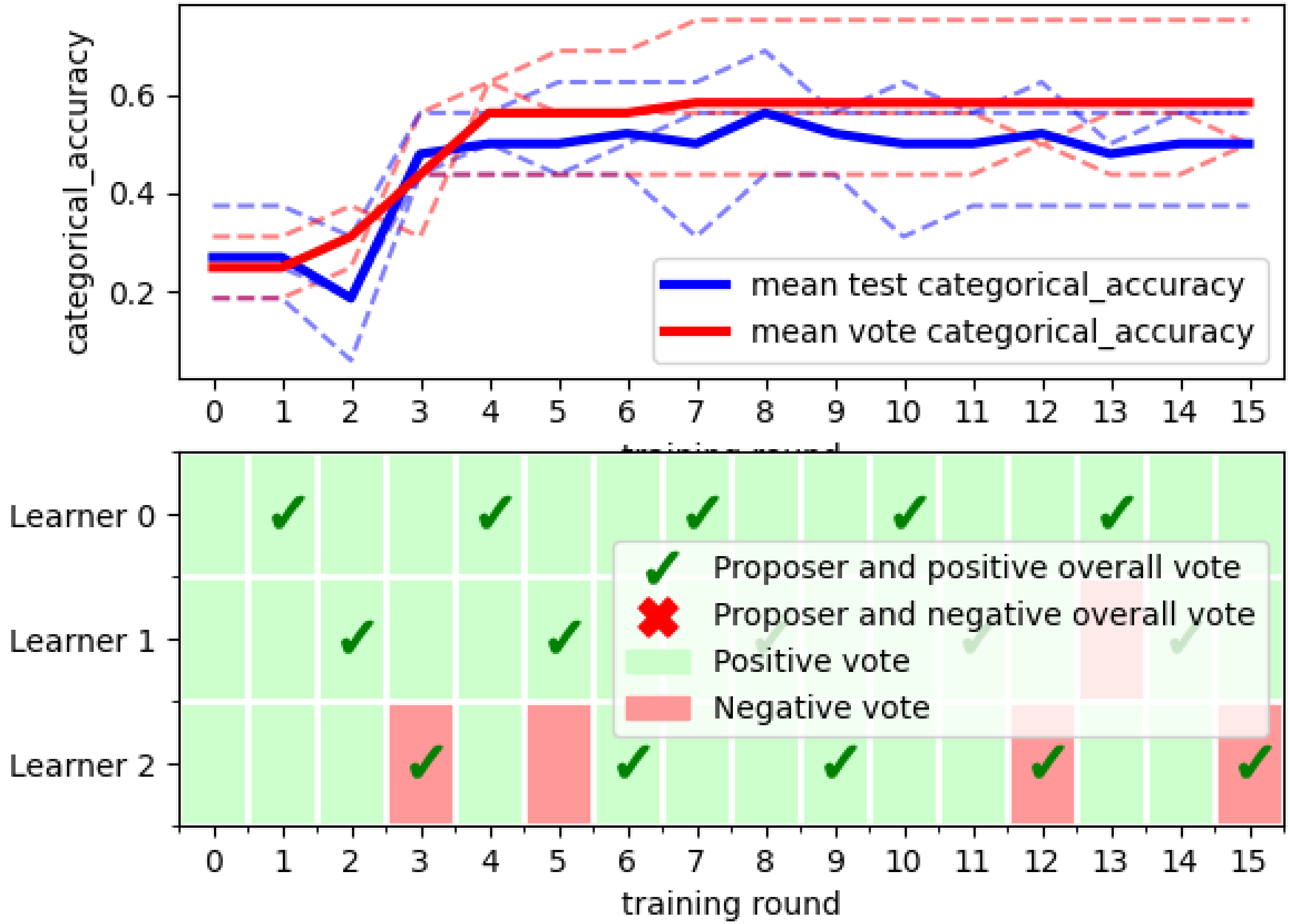
0. NN
1. ν_μCC
2. ν_eCC



DenseNet-169

Validation accuracy **68.6%**

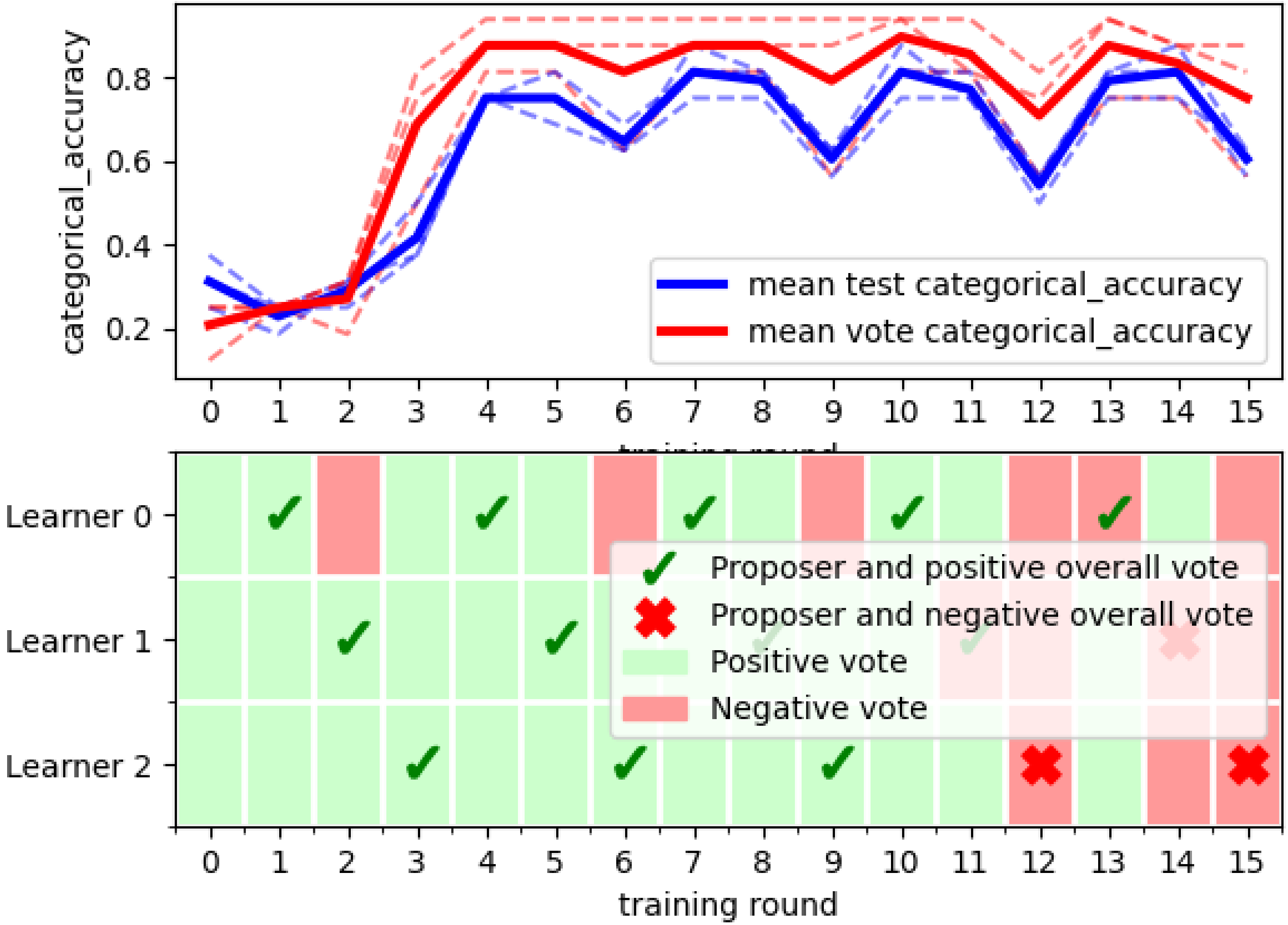
Results – ResNet 50 V2



3 Learners

15 Rounds

Results – DenseNet 169



3 Learners
15 Rounds

Results

- Efficiency tends to ramp up and reaches a plateau during the first 4 rounds.
- The more rounds we add, the higher is the possibility that a new set of weights is not chosen.
- Bad datasets tend to be ignored already during the first rounds.
- Mean vote tends to have higher efficiency than mean test.
- The overall efficiency is similar to the one obtained training the model with a single dataset.
- After reaching the plateau, the efficiency starts to oscillate.

Conclusion and Future Work

- ✓ Exponential usage of AI in particle physics requires new technologies to compare different ML models and datasets.
- ✓ This experiment worked and showed that this library can be successfully used also for very complicated datasets, as the ones used in particle physics.

- ➢ Several tests on scalability and benchmarks with different Keras models are now needed.
- ➢ Implementation on a new approach to test different ML models on the same dataset on planning stage.
- ➢ Colearn and in general Collaborative Learning are ready to be used by the community!

Thank you for your attention!!

For any questions:
sv408@hep.phy.cam.ac.uk