FERMILAB-SLIDES-18-112-E

USING MACHINE LEARNING TECHNIQUES FOR DATA QUALITY MONITORING AT CMS EXPERIMENT

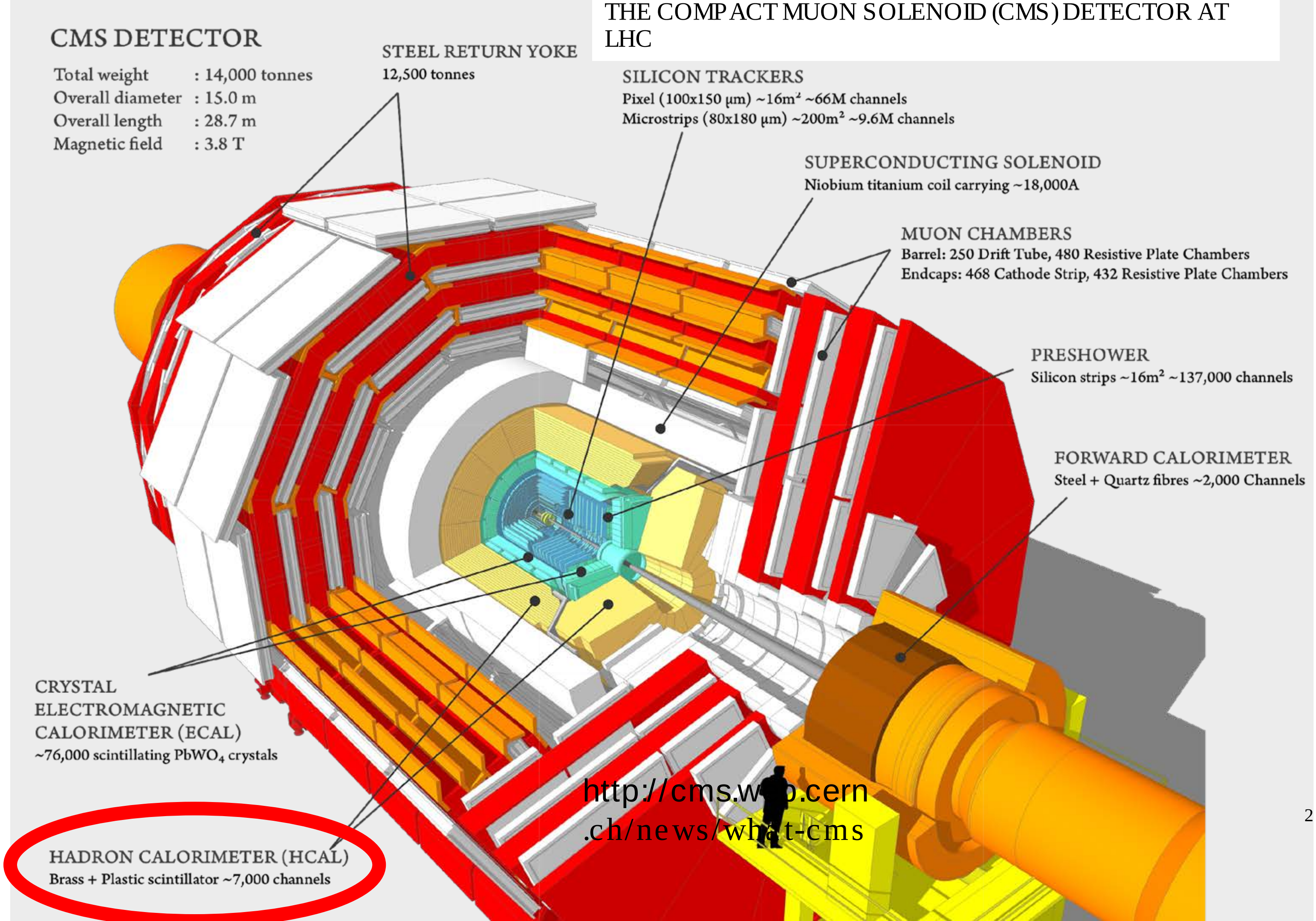
GUILLERMO A. FIDALGO RODRÍGUEZ

PHYSICS DEPARTMENT

UNIVERSITY OF PUERTO RICO MAYAGÜEZ

This document was prepared by [CMS Collaboration] using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359.

THE COMPACT MUON SOLENOID (CMS) DETECTOR AT LHC



http://cms.web.cern.ch/news/what-cms

OBJECTIVES

- Apply recent progress in Machine Learning techniques regarding automation of DQM scrutiny for HCAL
  - To focus on the Online DQM.
  - To compare the performance of different ML algorithms.
  - To compare fully supervised vs semi-supervised approach.
- Impact the current workflow, make it more efficient and can guarantee that the data is useful for physics analysis.

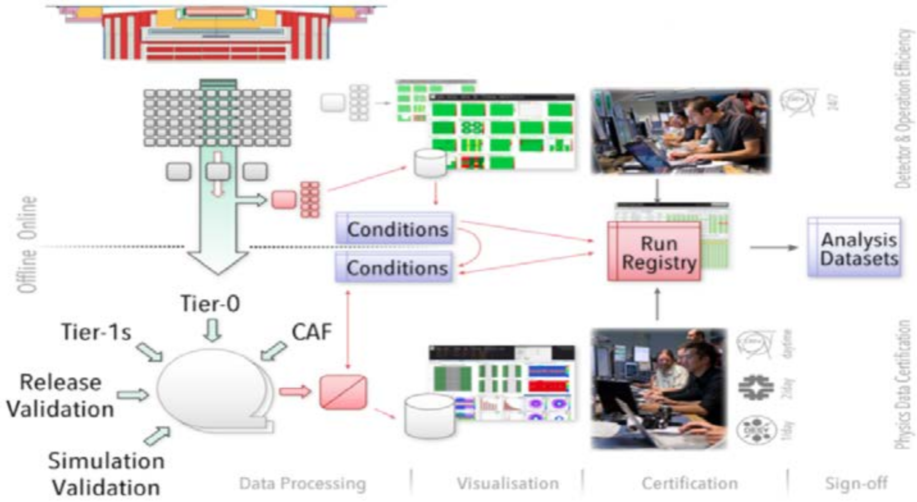
CHALLENGE

- Make sure detector behaves well to perform sensible data analysis.
- Reduce man power to discriminate good and bad data, spot problems, save time examining hundreds of histograms.
  - By building intelligence to analyze data, raise alarms, quick feedback.

- Implementing the best architecture for neural networks
  - Underfitting - Too simple and not able to learn
  - Overfitting - Too complex and learns very specific and/or unnecessary features
- There is no rule of thumb
  - Many, many, many……possible combinations.

WHAT IS DATA QUALITY MONITORING (DQM)?

- Two kinds of workflows:
- Online DQM
  - Provides feedback of live data taking.
  - Alarms if something goes wrong.
- Offline DQM
  - After data taking
  - Responsible for bookkeeping and certifying the final data with fine time granularity.



HYPOTHESIS AND PROJECT QUERIES

Queries

- Can we make an algorithm that identifies anomalies in the data flow?

Hypothesis

- We can develop a ML algorithm that takes the images as data and determine whether or not an error is occurring.

Rationale

- Since this algorithm takes images as inputs it can learn to compare the images given with a baseline and correctly identify patterns and deviations from the baseline.

TOOLS AND DATA PROCESSING

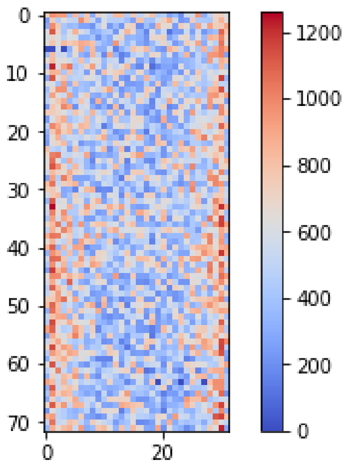
- Working env: python Jupyter notebook
- Keras (with Tensorflow as backend) and Scikit-learn
  - Creation of a model
  - Train and test its performance
- The input data consists of occupancy maps
  - one map for each luminosity section
  - Used 2017 good data and generate bad data artificially

IMAGE ANALYSIS TERMINOLOGY

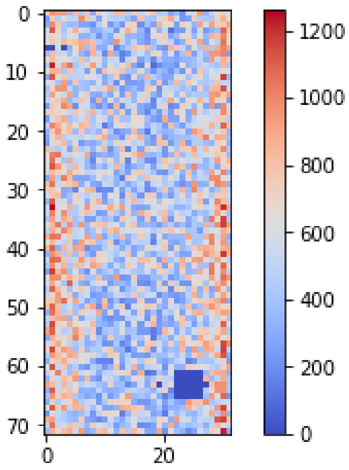
- Hot - image with noisy (red) channels
- Dead - image with inactive (blue) channels
- Good - regular images that are certified for analysis
- Model - an ML algorithm's structure
- Loss - number that represents distance from target value

IMAGES AND READOUT CHANNELS USED AS INPUTS FOR THE ML ALGORITHM

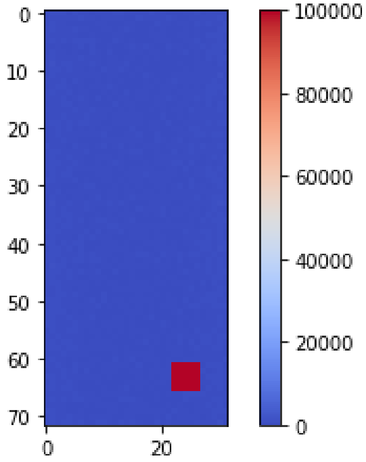
- Supervised and Semi-Supervised Learning
- 5x5 problematic region with random location
- 5x5 (readout channels) problematic region with fixed location



Good



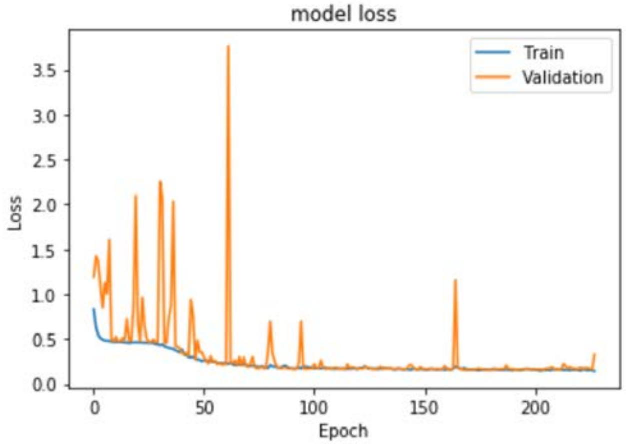
Dead

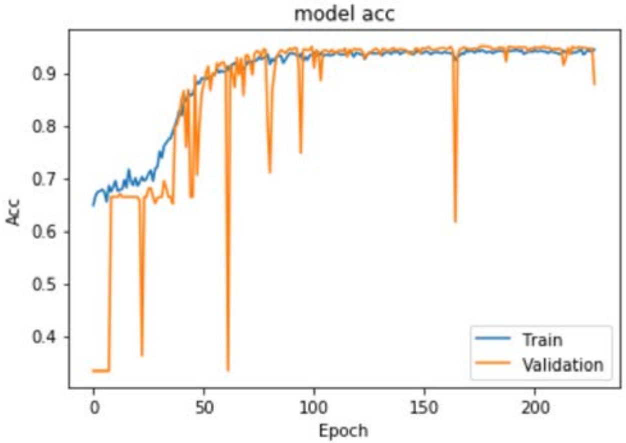


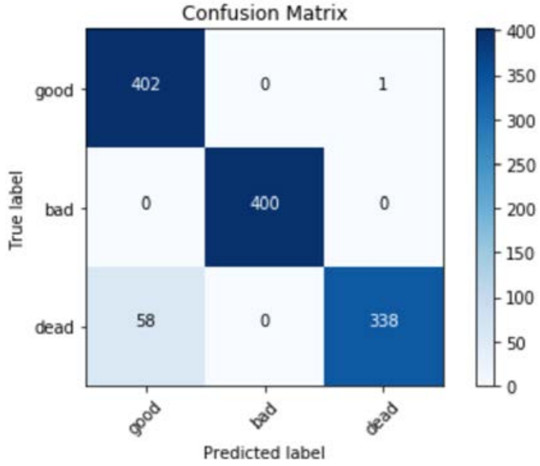
Hot

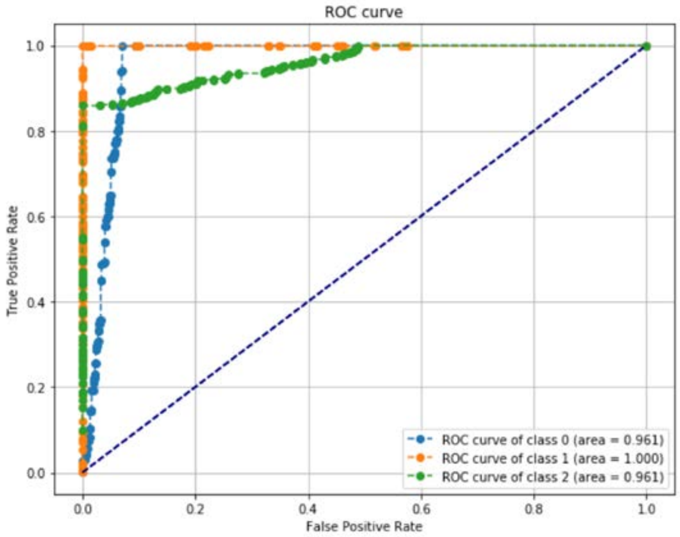
SUPERVISED LEARNING

accuracy score: 0.950792326939

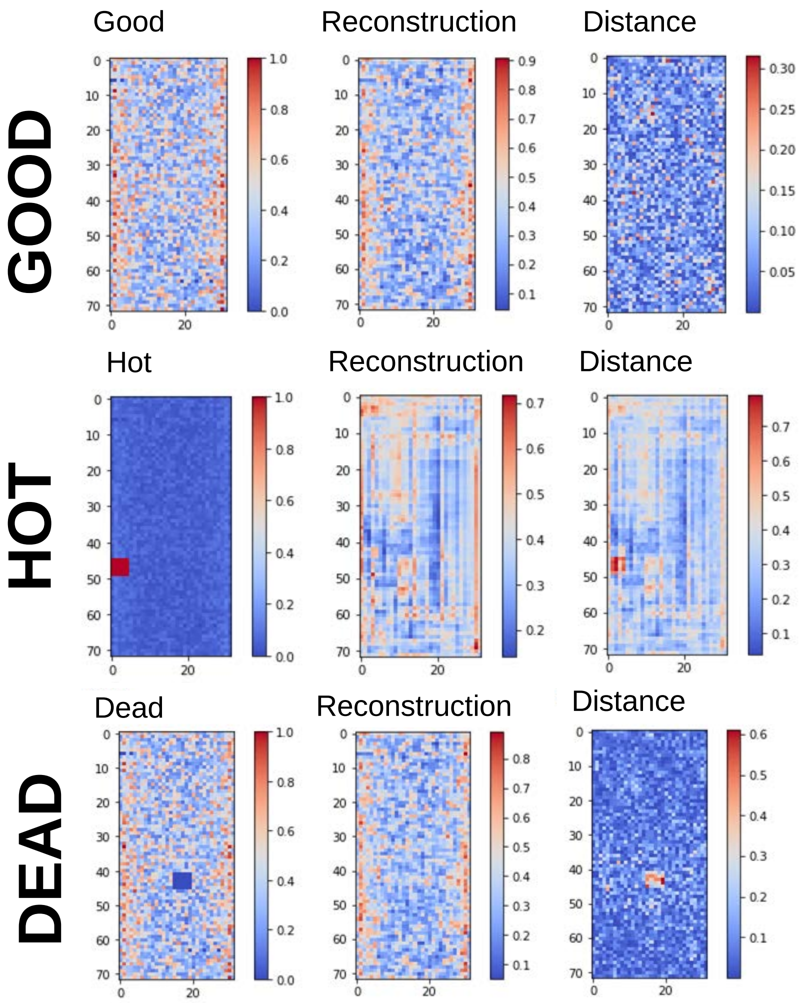






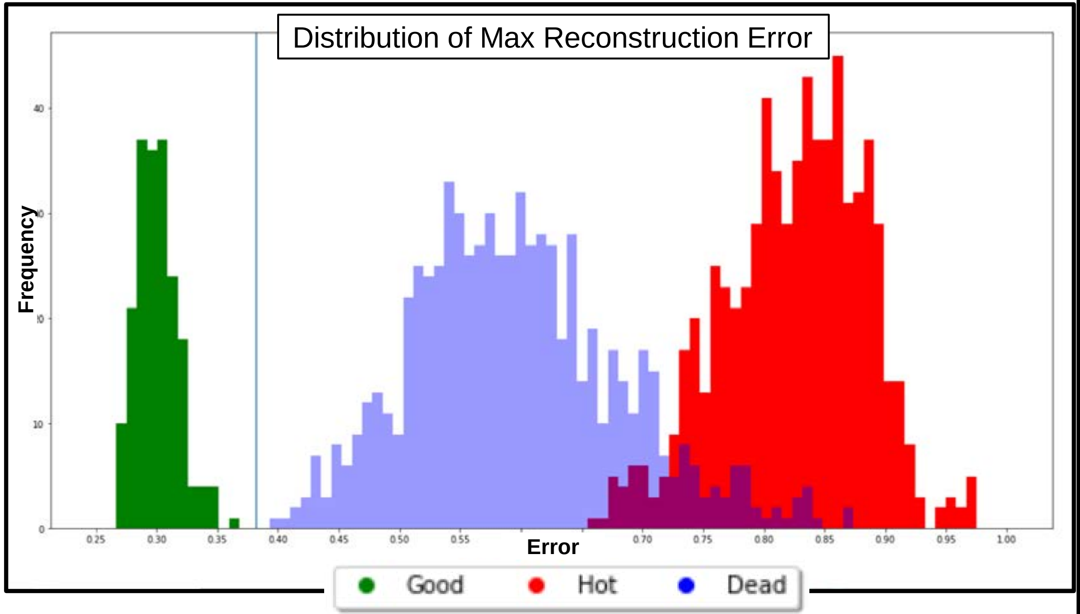


SEMI SUPERVISED LEARNING



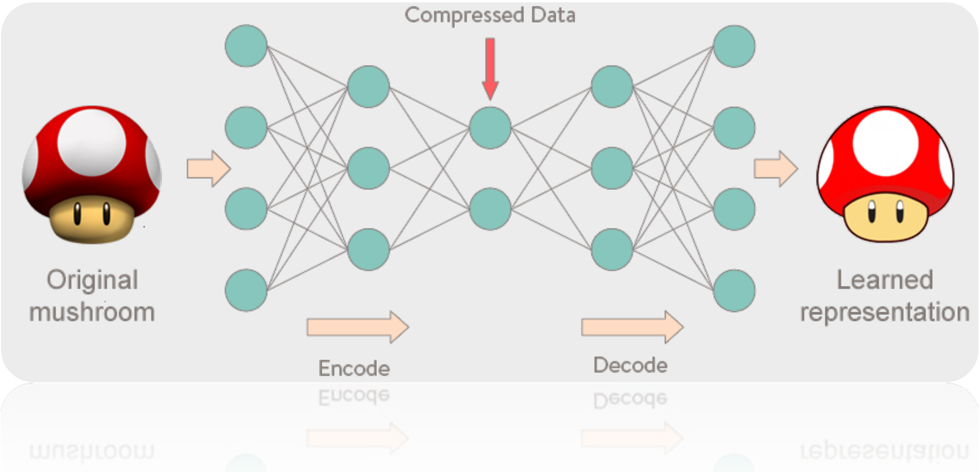
- Trained only on good images
- Expected to see better reconstruction for good images and a much different reconstruction for bad images.
- Bad images have 5x5 bad regions
  - Hot
  - Dead
- Images have been normalized
- this architecture seems to perform best for us.

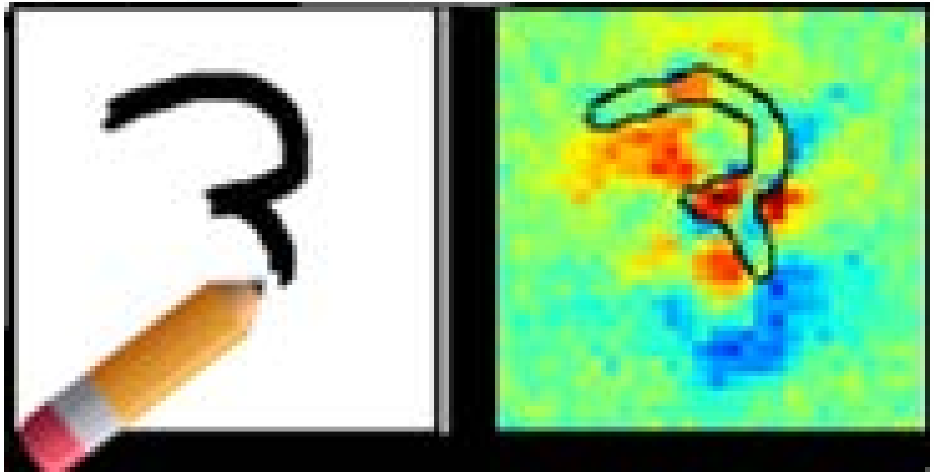
ERROR DISTRIBUTION PER IMAGE CLASS



WHAT'S NEXT?

- Why and exactly what is it learning?
- Can we make it work with something more realistic?
  - 1x1 bad region (channel)
  - Can it identify what values should be expected after each lumi-section?
  - Move from artificial bad data to real cases of bad data (in progress)





Acknowledgments

- The US State Dept.
- The University of Michigan
- CERN/CMS
- Federico De Guio , Ph.D (Texas Tech)
- Nural Akchurin, Ph.D (Texas Tech)
- Sudhir Malik , Ph.D (University of Puerto Rico Mayagüez)
- Steven Goldfarb, Ph.D (University of Melbourne)
- Jean Krisch, Ph.D (University of Michigan)

Thank You!