# Anomaly Detection Using Deep Sparse Autoencoders for CERN Particle Detector Data

**Filip Široký <fsiroky@mail.muni.cz>**
MIR research group <mir.fi.muni.cz>, Faculty of Informatics, Masaryk University, Brno, Czechia

## Introduction

The certification of Compact Muon Solenoid (CMS) data, as usable for physics analysis, is crucial to ensure the quality of all results published by CERN. The certification conducted by human experts is time-consuming and causes data loss.

Our goal is to automate this process, discarding bad lumisections (i.e. 23 seconds of data taking) instead of entire runs (i.e. 400 times more data on average), as well as to provide decent level of interpretability of the model's decisions.

## Data representation and preprocessing

Human experts make decisions regarding the data quality based on histograms. We decided to represent each sample as a 2807-dimensional vector that is composed of five quantiles, the mean and the standard deviation of all collection distributions (photons, muons, etc.). Therefore, every part of the detector is well-represented.

## Autoencoder architecture

We use the autoencoder as a semi-supervised method rather than as an unsupervised method for dimensionality reduction. As a semi-supervised method, the autoencoder addresses all our present issues: class imbalance, sparsity of anomalies as they occur due to various reasons, changing configuration of both the Large Hadron Collider (LHC) and CMS in time, and the curse of dimensionality.

We train the autoencoder on good lumisections only, teaching it to reconstruct well only those new lumisections that are similar to the good training data, and reconstruct poorly all the rest (based on some metric, e.g. the Mean Squared Error, MSE). We can then set a threshold on the reconstruction error and use the model as a binary classifier.

We introduced regularizers such as L1 to avoid overfitting by penalizing large coefficients. This made the model sparser and increased performance.

## Model interpretability

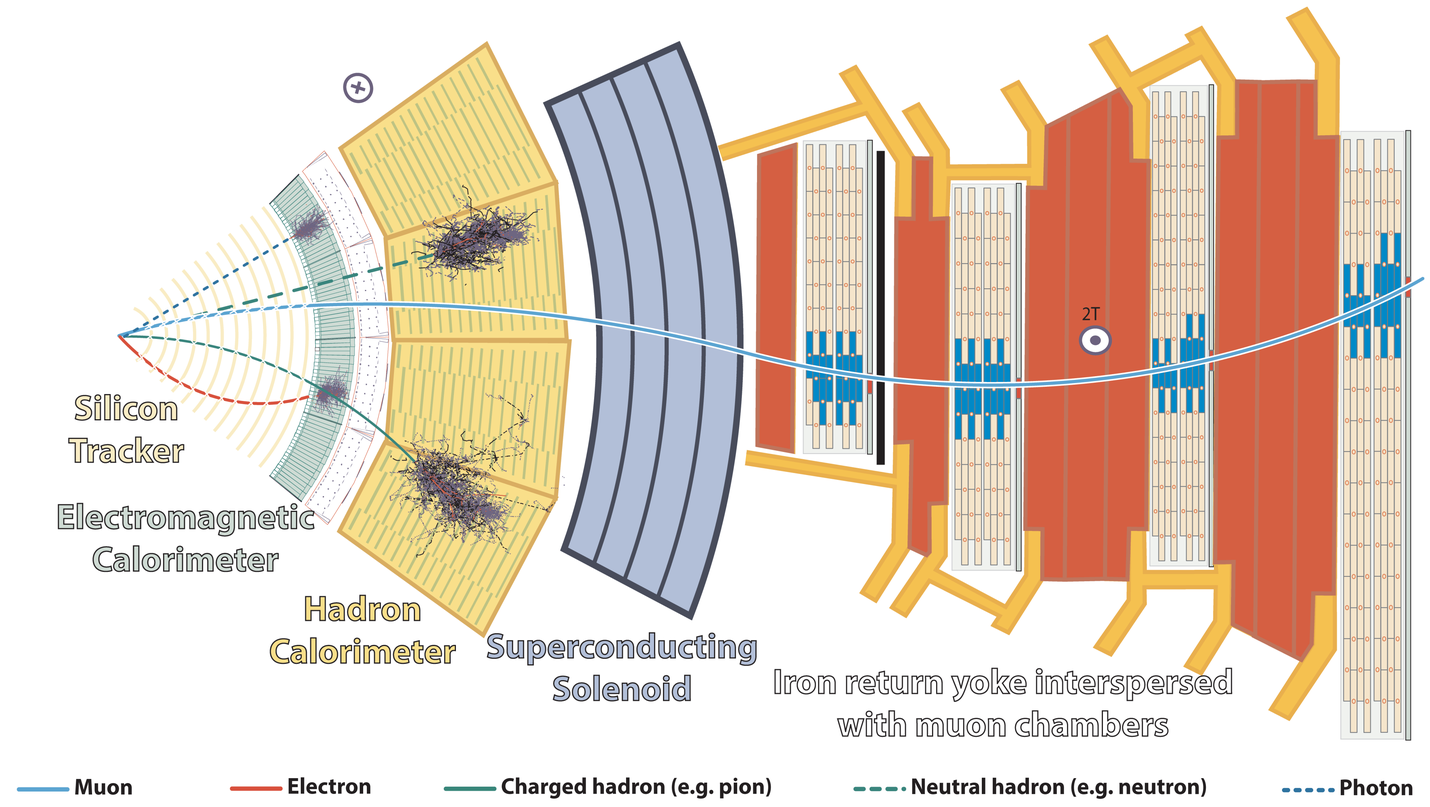
For any lumisection in the test set, we can plot the difference between the actual and reconstructed values for all the features of that lumisection. We can then group the features based on which physics object they represent and plot a reconstruction error distribution. We would expect to see low uniform reconstruction error for good lumisections and high average error for anomalies with peaks in the features corresponding to the objects it reconstructed poorly. These peaks suggest there was a problem with particular subdetector, e.g. a peak at muon-related features would suggest a problem with the muon chambers.

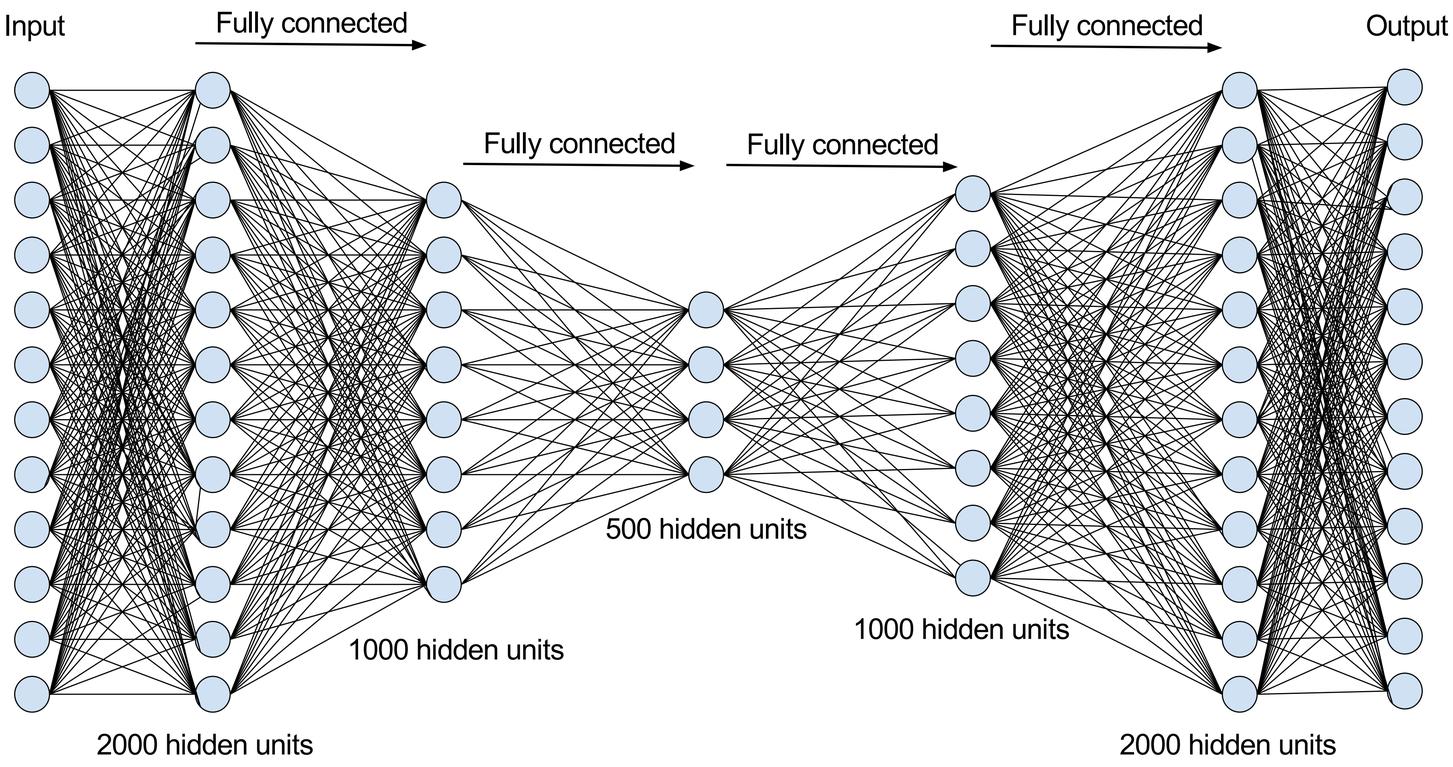
## Conclusion

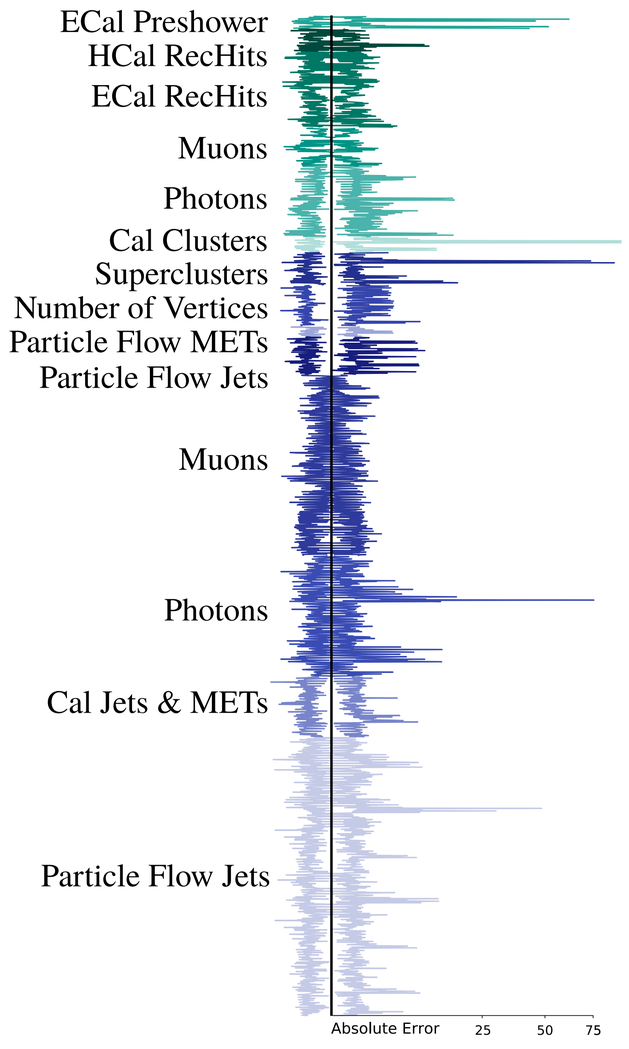
We have demonstrated that our autoencoder is a useful semi-supervised method (ROC AUC ≈ 0.97) that performs comparably to the state of the art and is much more interpretable. Our project will continue to integrate with the automated data processing infrastructure of CMS, saving expensive data and manmonths of labour every year.

## Bibliography

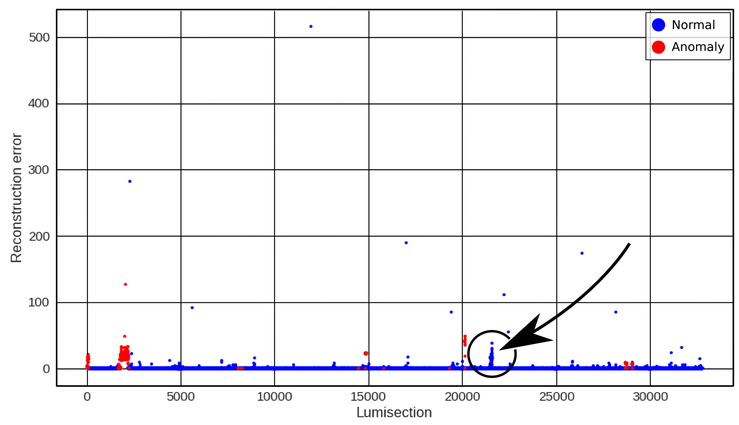
POL, Alan A. et al. Anomaly detection using Deep Autoencoders for the assessment of the quality of the data acquired by the CMS experiment. Presented at Computing in HE Physics. 2018.



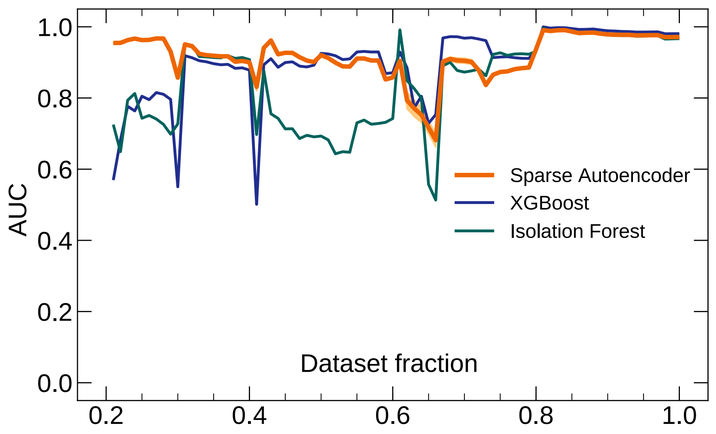




Reconstruction error for features of a random good lumisection (left) and a bad lumisection indicating an Electromagnetic Calorimeter problem (right).




Lumisection reconstruction error MSE over all features. The arrow shows a group of lumisections that were originally labeled as good, but were later discovered to be anomalous thanks to our results.




Our autoencoder achieves comparable performance (ROC AUC) to state-of-the-art classifiers.