Anomaly Detection Based on Machine Learning for the CMS Electromagnetic Calorimeter Online Data Quality Monitoring

A. Harilal, K. Park, Manfred Paulini (Carnegie Mellon University)

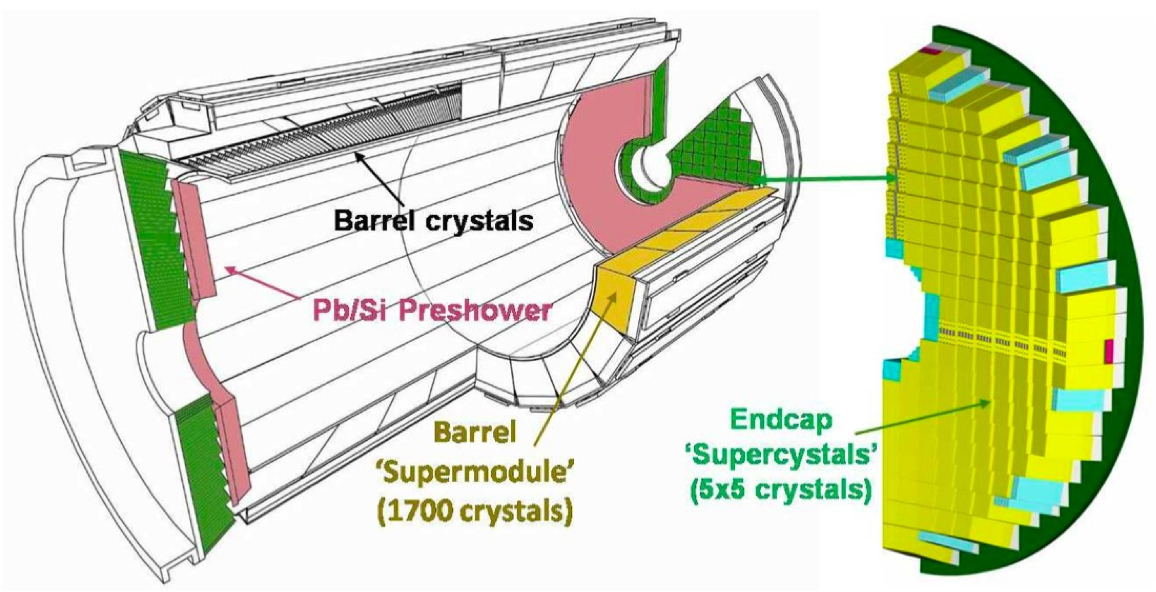
On behalf of the CMS Collaboration

Summary

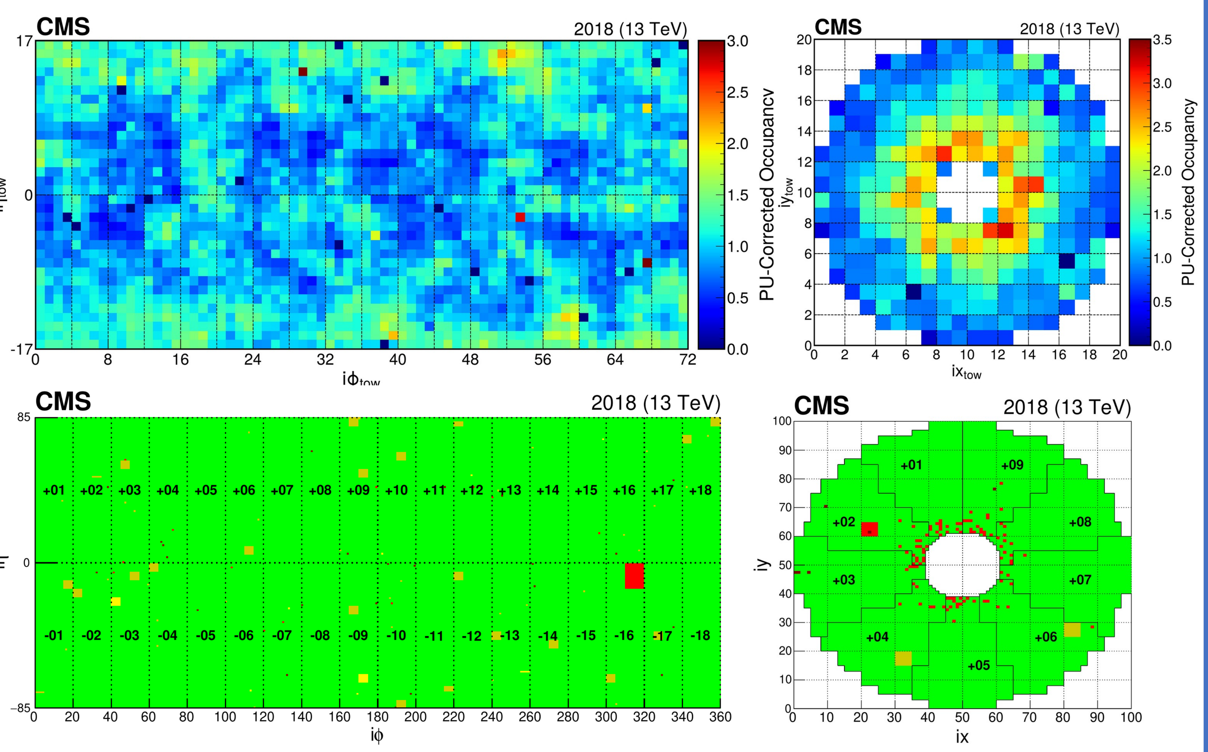
- Online Data Quality Monitoring (DQM) of CMS electromagnetic calorimeter (ECAL) is vital operational tool
- Allows detector experts to quickly identify and diagnose broad range of detector issues that could affect quality of physics data
- Developed real-time autoencoder (AE) based anomaly detection system using semi-supervised machine learning (ML) enabling detection of anomalies
- Novel application of spatial and time corrections yields order of magnitude improvement in AE performance for 99% anomaly tagging rate
- Validations on real anomalies from CMS data shows AE-based system is able to spot anomalies at tower-level (set of 5x5 ECAL crystals) granularity
- Deployment of AE-based system in CMS DQM workflow for LHC Run3 shows system performs well in detecting anomalies and identifying degrading channels missed previously
- AE-based DQM system complements and strengthens existing DQM for ECAL at CMS

Introduction

CMS Electromagnetic Calorimeter

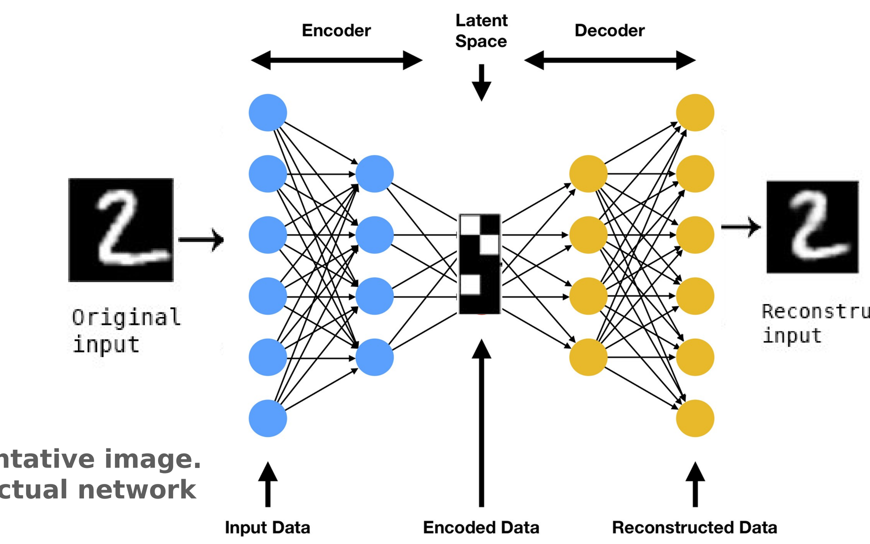


ECAL Data Quality Monitoring

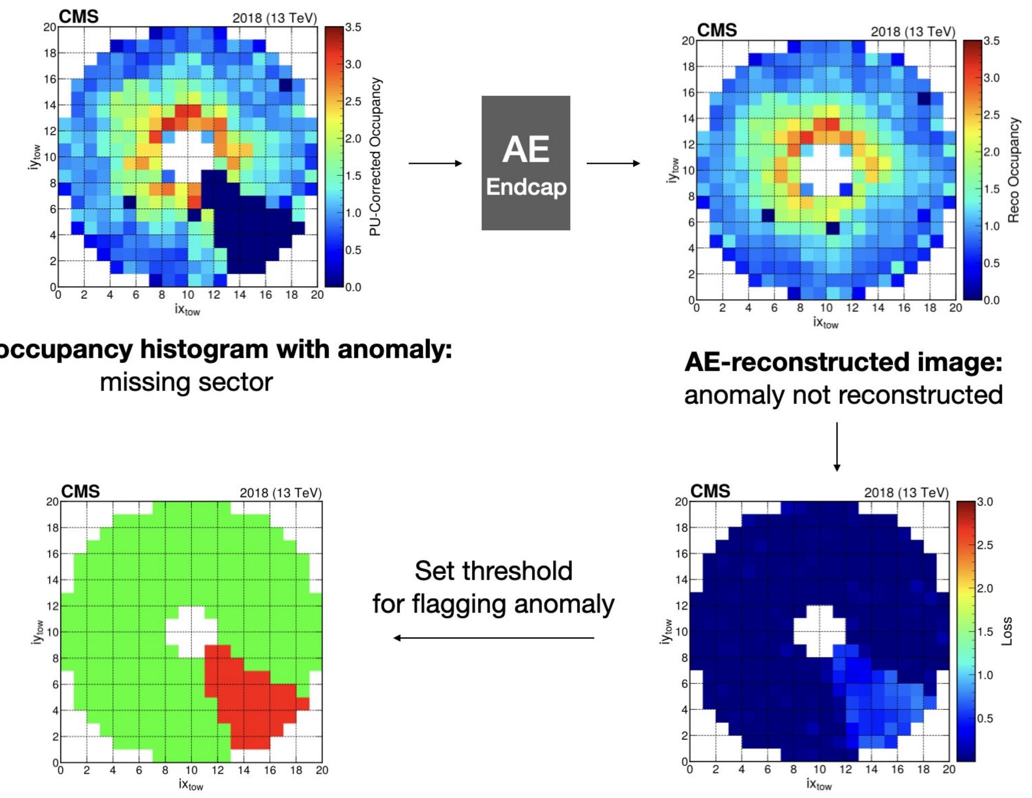


Methodology

ML Based Anomaly Detection with Autoencoder



(Representative image. Not the actual network used.)



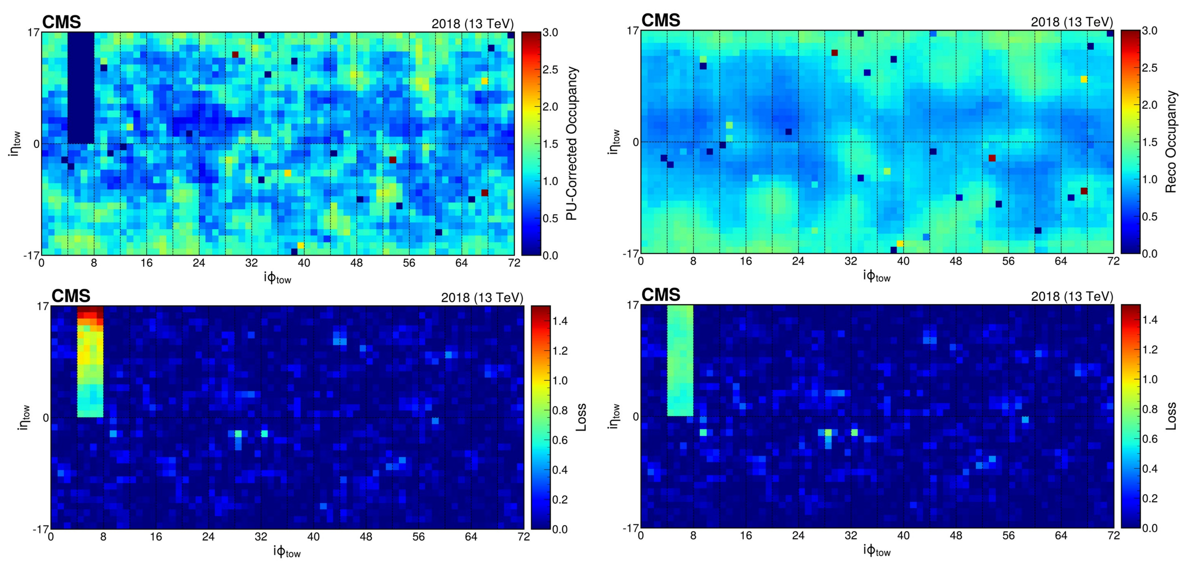
Input occupancy histogram with anomaly: missing sector

AE-reconstructed image: anomaly not reconstructed

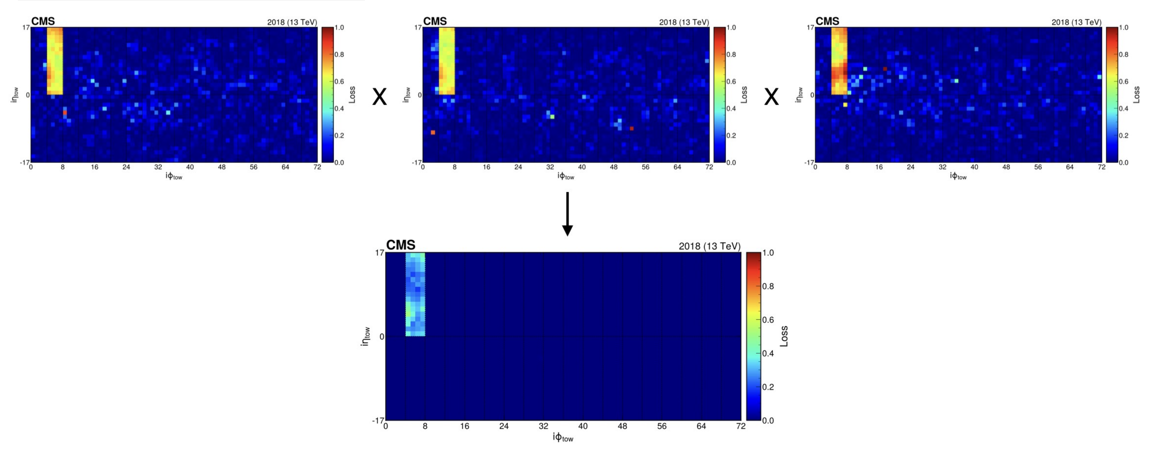
Final quality output: anomalous towers: red good towers: green

Loss map: anomalous showing high loss

Spatial Response Correction



Time Correction

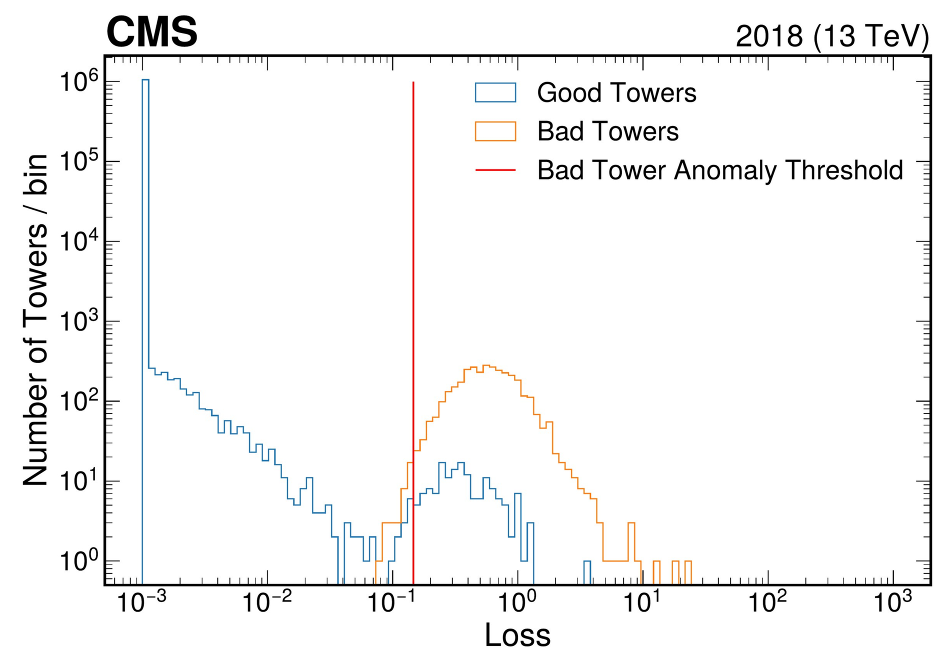


Results

Performance Metric

False Discovery Rate (FDR)

FDR = #good towers over threshold / #good & bad towers

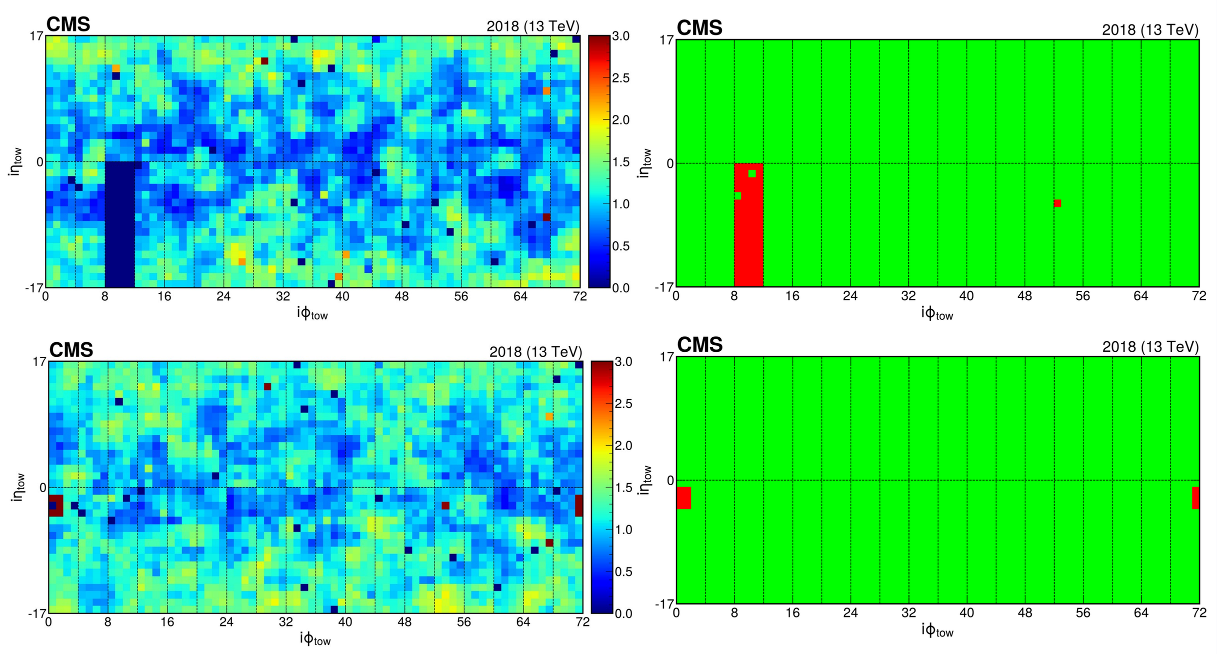


Testing on Fake Anomaly Scenarios

	FDR for 99% anomaly detection		
	Missing Supermodule	Zero Occupancy Tower	Hot Tower
Baseline no correction	14%	90%	5.2%
Baseline after time correction	5.9%	80%	$< 0.01\%$
AE no correction	3.6%	51%	2.8%
AE after spatial correction	3.1%	49%	2.9%
AE after spatial and time corrections	0.13%	4.1%	$< 0.01\%$

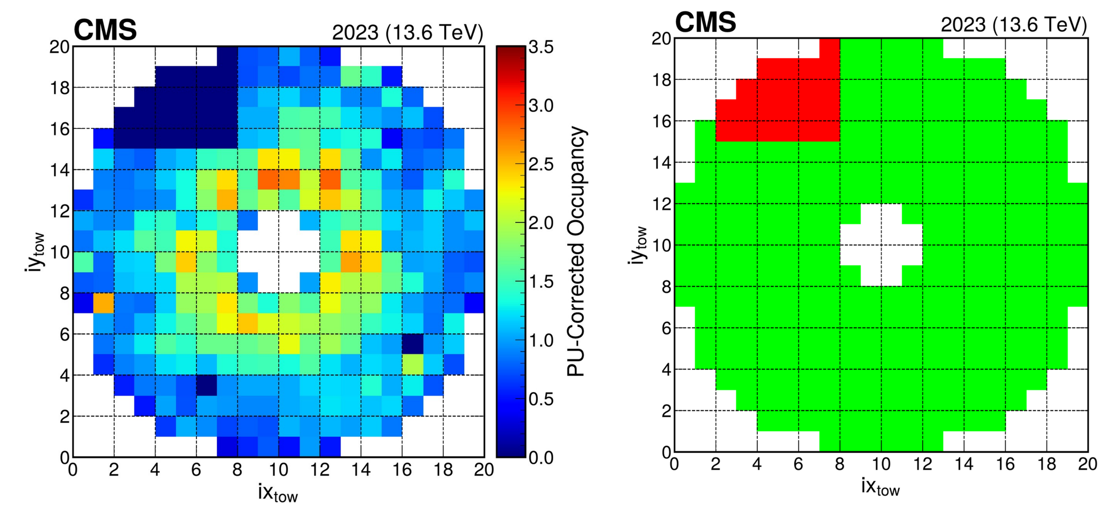
	FDR for 99% anomaly detection					
	Missing Sector		Zero Occupancy Tower		Hot Tower	
	EE+	EE−	EE+	EE−	EE+	EE−
AE no correction	29%	28%	86%	86%	$< 0.01\%$	$< 0.01\%$
AE after spatial correction	1.8%	2.2%	11%	14%	0.02%	0.04%
AE after spatial and time corrections	0.06%	0.18%	1.4%	4.4%	$< 0.01\%$	$< 0.01\%$

Testing on Real Anomalies

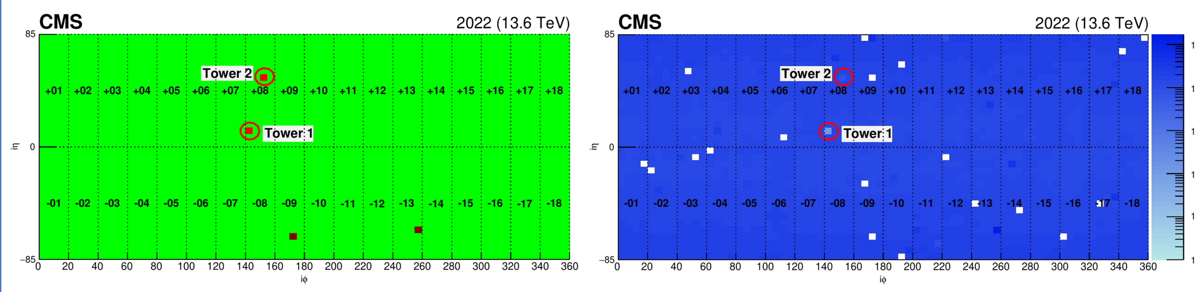


Performance during LHC Run 3

Deployed online between 2022 and 2023



Detecting New Degrading Towers



Conclusion

- Autoencoder based anomaly detection system using semi-supervised machine learning was developed for Online Data Quality Monitoring of CMS electromagnetic calorimeter
- Enables detection of detector anomalies in real time
- Novel application of spatial and time corrections yields strong performance for 99% anomaly tagging rate
- Validations on real anomalies from CMS data and deployment during Run 3 shows AE-based system is able to spot anomalies of various shapes, sizes, and locations at tower-level granularity using single threshold
- System can be generalized not only to other subsystems of the CMS detector but also to other particle physics experiments for anomaly detection and data quality monitoring

Acknowledgments

More information **arXiv:2309.10157 [physics.ins-det]**

Work accepted for publication in *Comput. Softw. Big Sci.*

We congratulate our colleagues in the CERN accelerator departments for the excellent performance of the LHC and thank the technical and administrative staffs at CERN and at other CMS institutes for their contributions to the success of the CMS effort. In addition, we gratefully acknowledge the computing centres and personnel of the Worldwide LHC Computing Grid and other centres for delivering so effectively the computing infrastructure essential to our analyses. Finally, we acknowledge the enduring support for the construction and operation of the LHC, the CMS detector, and the supporting computing infrastructure provided by many funding agencies.

This research is supported in part by the U.S. Department of Energy, Office of Science, through DOE award DE-SC0010118.