Anomaly detection using statistical and AI-based approaches

Dr. Balázs Harangi

Debrecen, November 28, 2025

DEBRECENI EGYETEM
INFORMATIKAI KAR

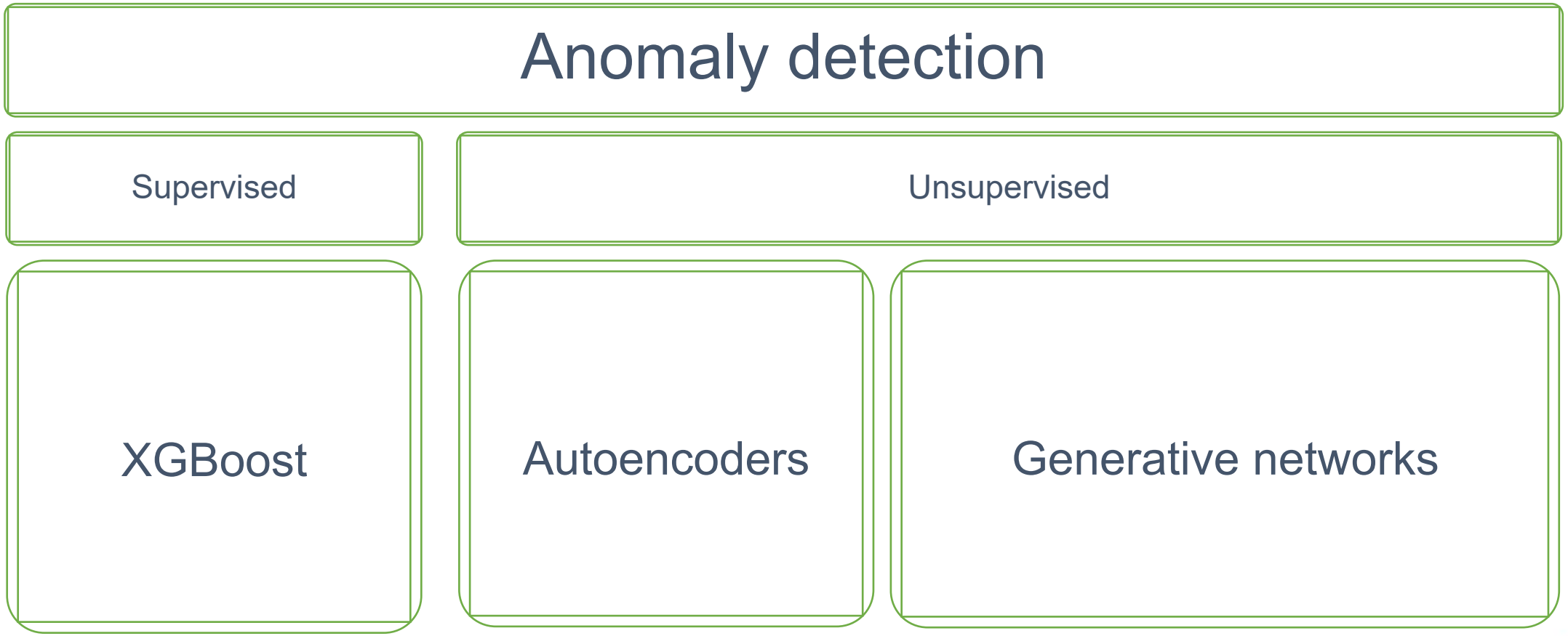
WHAT IS AN ANOMALY?

- A data point that differs significantly from other data points
- Anomaly detection can be useful in telecommunications/network systems, cybersecurity, finance, industry, IoT, healthcare, autonomous driving, video surveillance, and robotics.

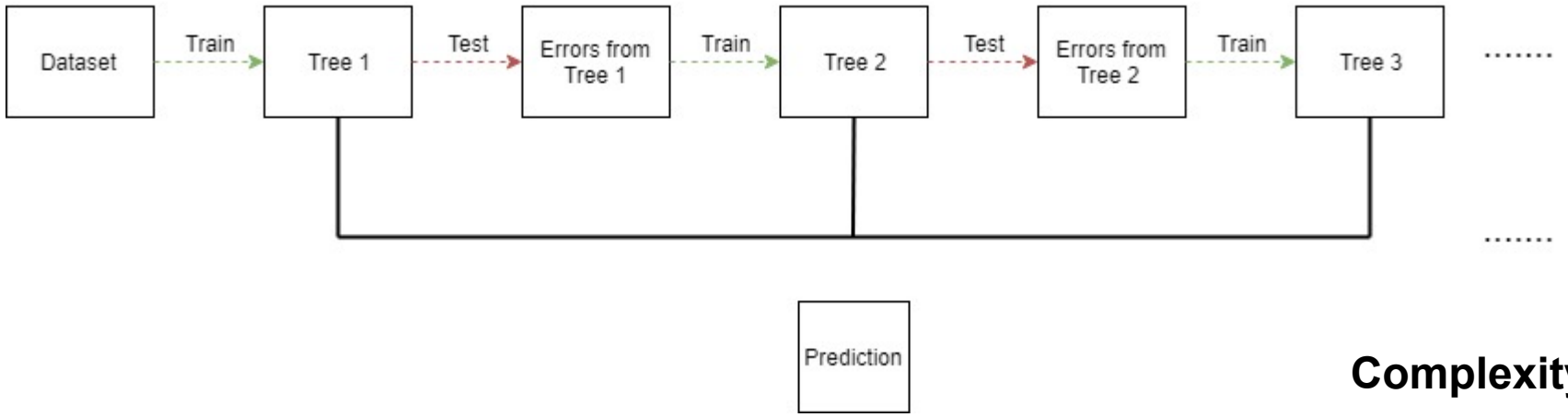
PRACTICE

- What are the scenarios that cause anomalies in your organization/domain?
- Which data sources can influence or record these anomalies?
- What data analysis techniques could be used or have already been used to detect these events?

DETECTION METHODS



XGBoost supervised learning solution



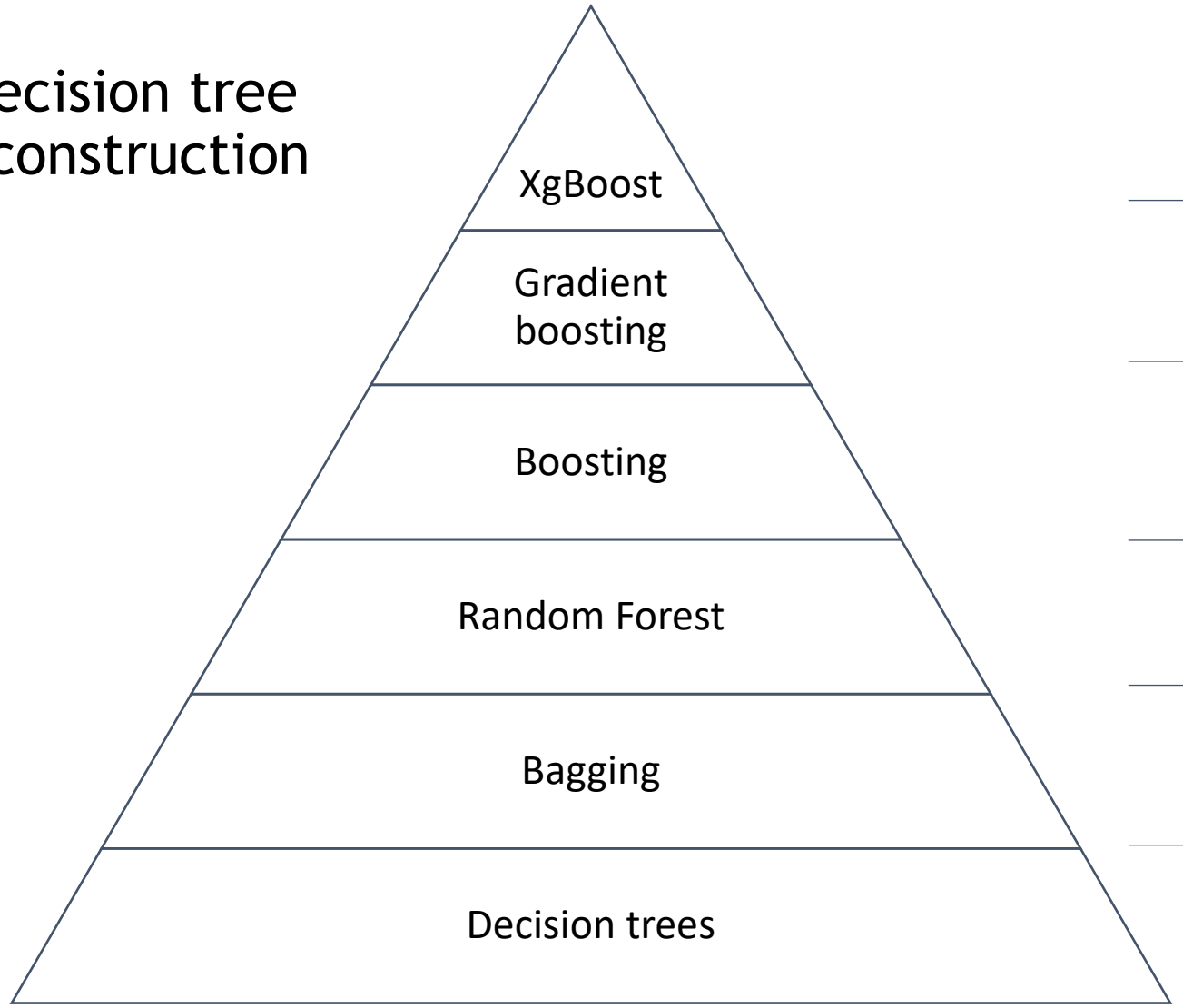
Build trees one at a time, where each new tree helps correct the mistakes made by the previously trained tree.

Complexity of trees

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Loss function

Decision tree construction



An optimized version of GBT that includes parallelism, tree pruning, and regularization.

Use of Gradient Descent to minimize the errors of trees built in sequence.

Trees that are built sequentially, minimizing the errors of previous trees and giving greater weight to those that perform better.

Use of random subsets of data sets to build multiple decision trees.

A set of decision trees that makes decisions by majority vote.

A tree-based algorithm that makes decisions based on certain conditions.

UNSUPERVISED METHODS

- Statistical methods assume that data can be modeled based on a given distribution.

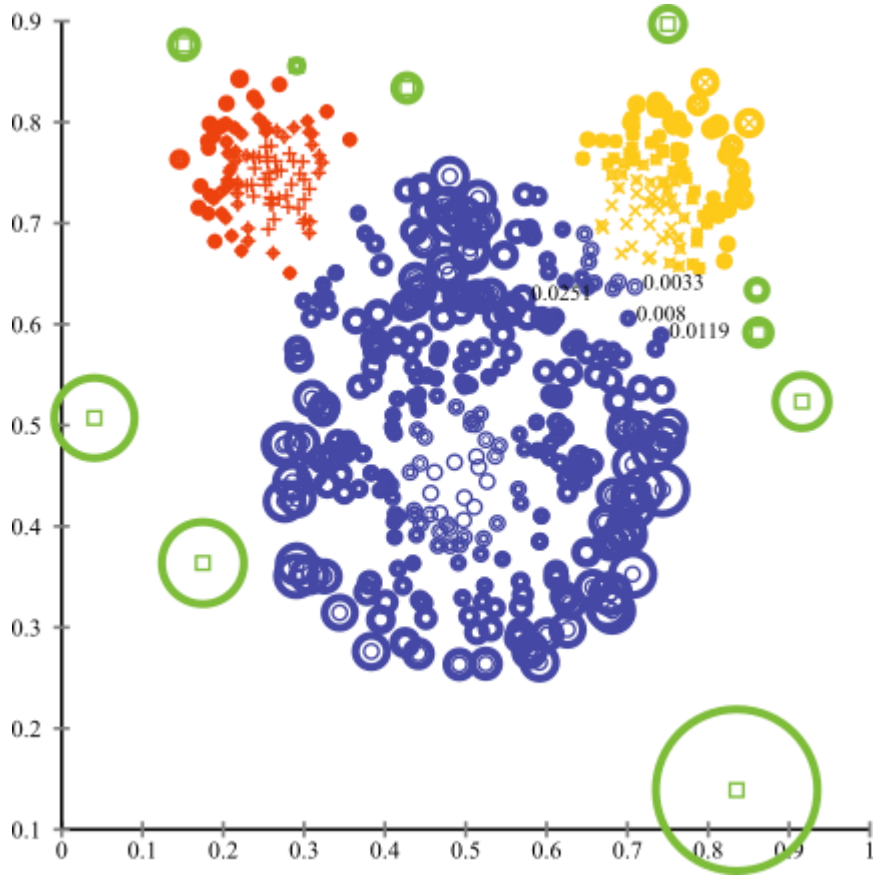
  An anomaly occurs when the probability is lower than the threshold value.

- Proximity methods use distance to determine anomalies

  An anomaly occurs when the distance from the centroid exceeds the threshold value

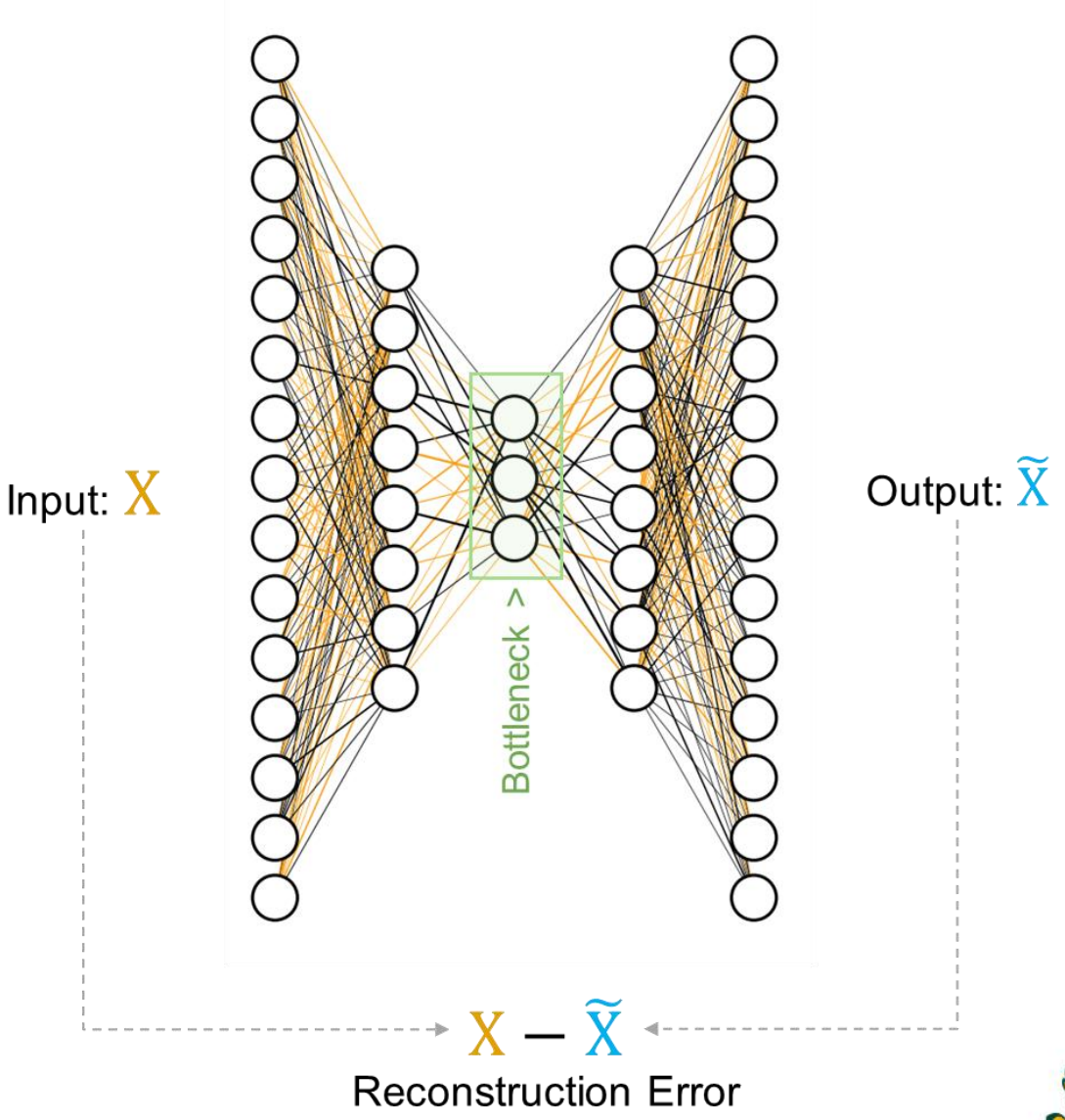
- Deviation methods use lower-dimensional embeddings and reconstruction errors

  An anomaly occurs when the reconstruction error is greater than one or the standard deviation



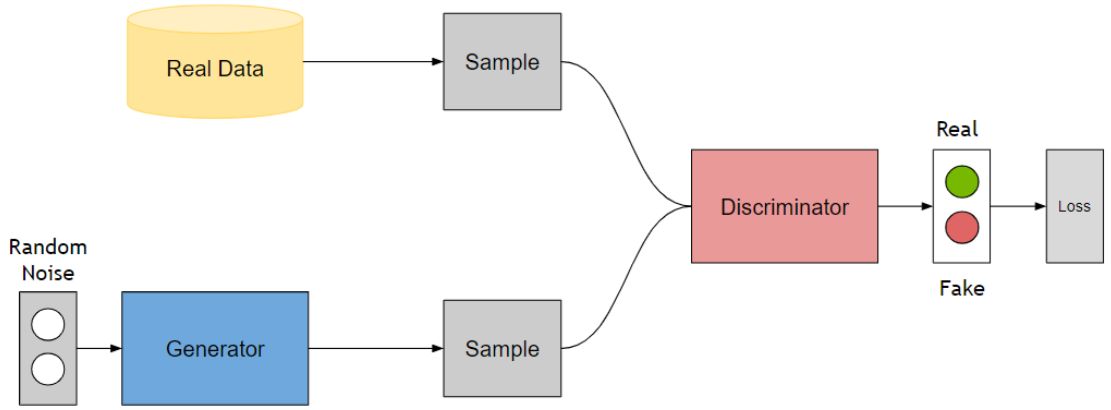
AUTOENCODERS

- Autoencoders are a type of unsupervised learning method that has applications beyond anomaly detection.
- An autoencoder consists of two parts: the encoder and the decoder.
- The encoder is a neural network that maps the input (typically) to a lower-dimensional space.
- The decoder is a neural network that remaps the encoded data back to the input.
- Anomalies are associated with large reconstruction errors.



GAN NETWORKS

- A generative model that learns to generate patterns with characteristics identical to those in the dataset.
- The generator, `G`, produces fake patterns.
- The discriminator, "D," receives samples from both G and the dataset.
- During training: the generator tries to deceive the discriminator by outputting values similar to real data, while the discriminator tries to improve its ability to distinguish between real and fake data.



High-energy physics and AI

- PHENIX detector
- direct photon analysis
- DHM@EMCal (25,000 towels)
- Strong energy dependence
- Patterns arising from HV problems

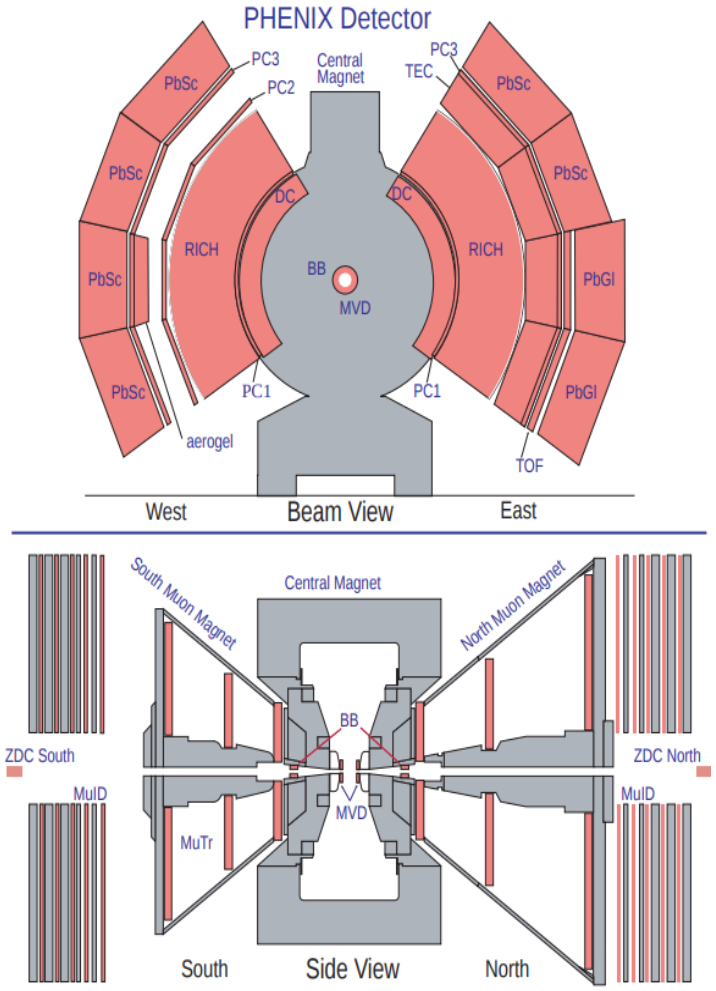
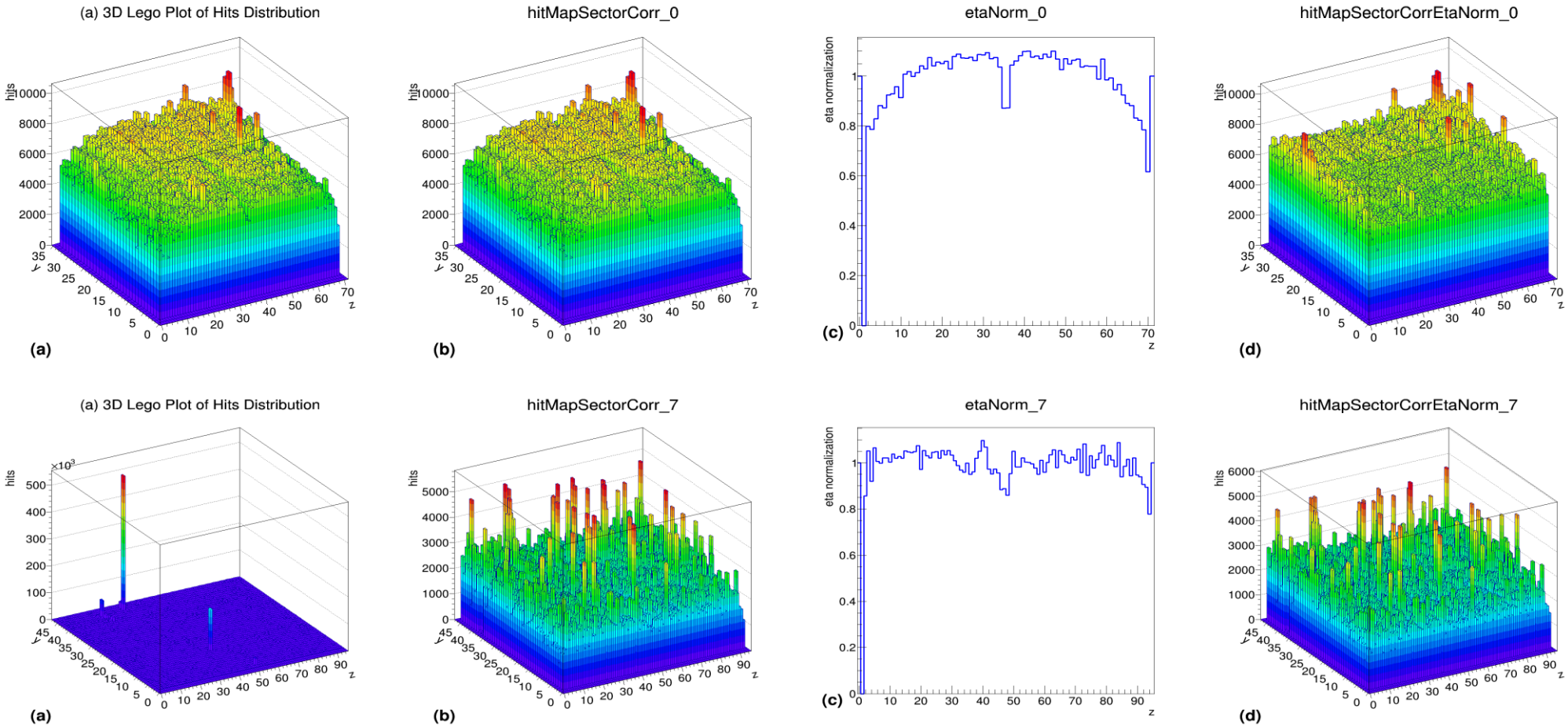


Figure 3.2: The PHENIX setup during the fourth RHIC beam period.

High-energy physics and AI

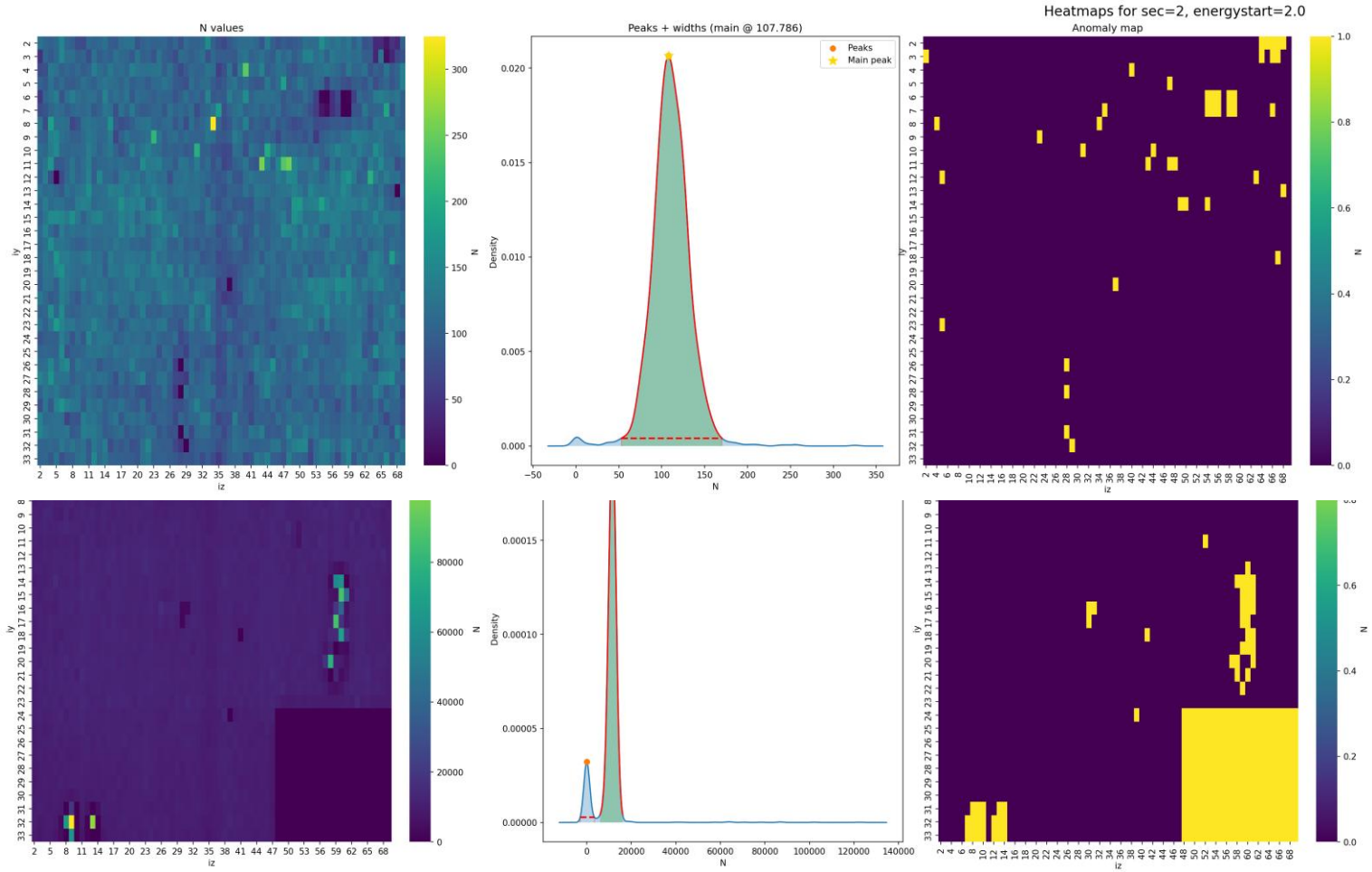


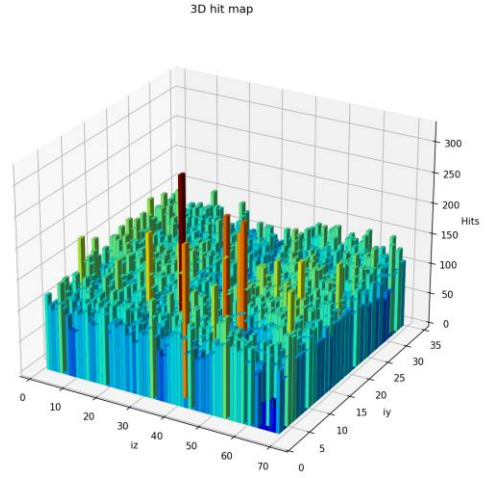
Debreceni Egyetem Informatikai Kar

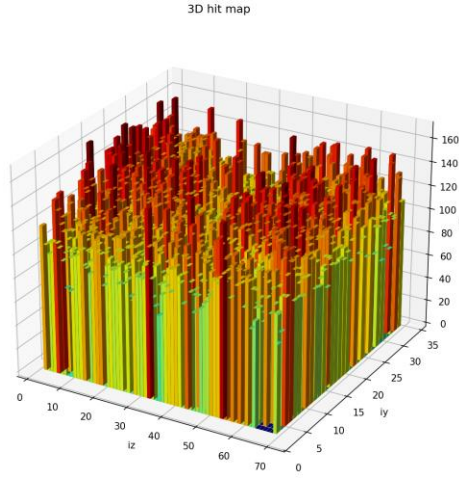
Application of statistical methods

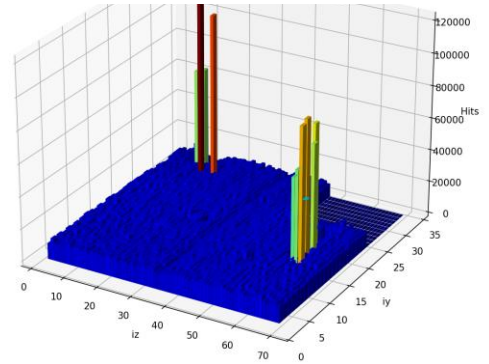
1. **Extracting the hit rate**
2. **Density estimation:** Examining the distribution of hits using kernel density estimation (KDE), which reveals how frequently values occur on the number line.
3. **Detection of peaks in the density function:** Identification of local maxima in the density curve, which represent frequent and typical value ranges (concentration regions).
4. **Selecting dominant peaks:** Retaining only statistically significant modes. These determine the normal operating ranges of the system. Currently, only the most dominant peak is used.
5. **Determining threshold values around the peak:** Determining threshold values by analyzing the decrease in the density curve. Extreme values beyond the peak boundaries represent anomaly zones.
6. **Anomaly detection and filtering**

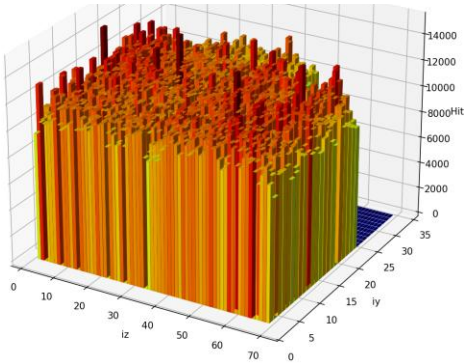
Application of statistical methods







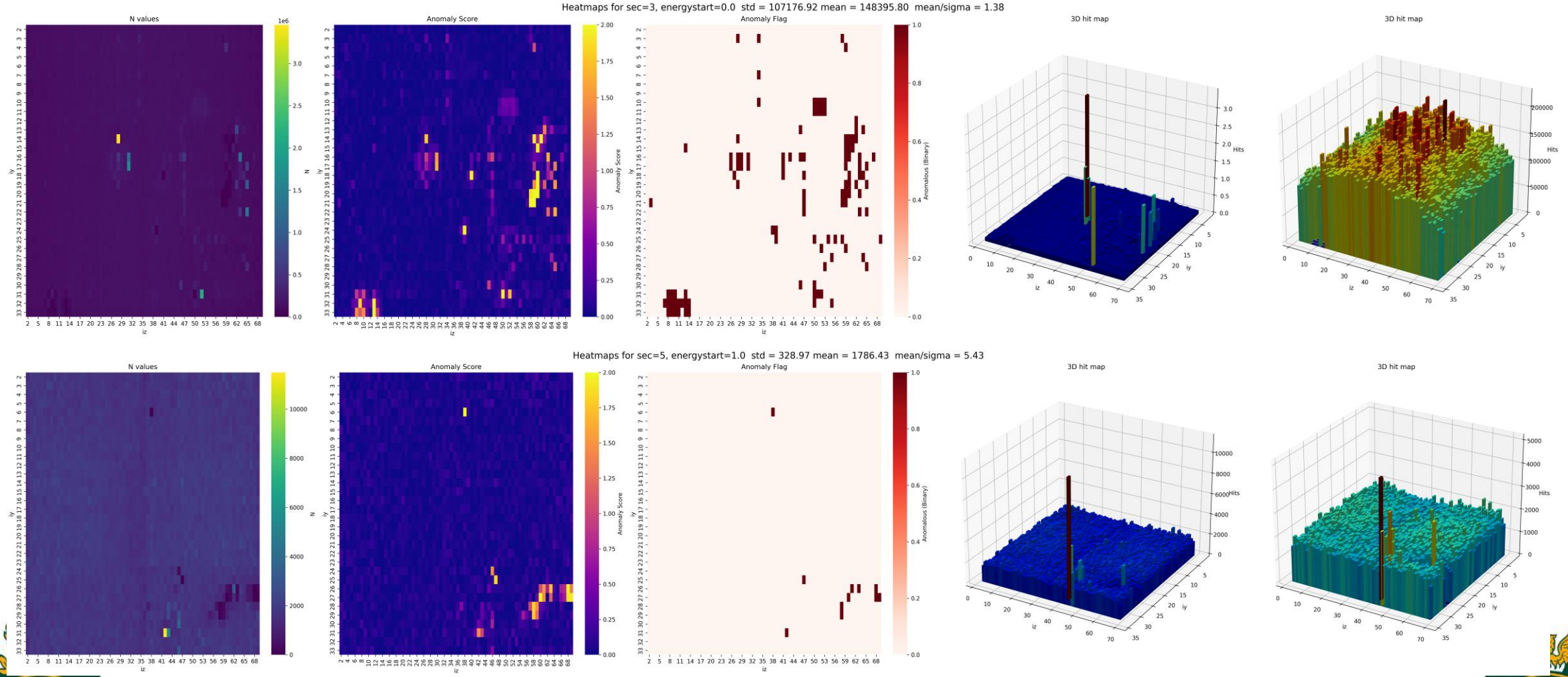




Anomaly detection using machine learning

1. **Extraction of statistical characteristics:**
   Calculation of descriptive statistics: percentiles (1%, 95%, 99%), median, mean, standard deviation, skewness, kurtosis, entropy, proportion of non-zero values.
2. **Feature selection:**
   Section, iy coordinate, iz coordinate, initial energy level, 99th percentile, median, 1st percentile, mean, standard deviation, skewness, kurtosis, entropy, proportion of non-zero values.
3. **Anomaly scoring using the SMAPE (symmetric mean absolute percentage error) indicator**
   Determining the degree of anomaly using SMAPE based on the difference between the estimated and actual hit rates.
4. **Anomaly detection with multiple decision thresholds**
   Application of mean and standard deviation-based thresholds to identify actual anomalies.
5. **Removing detected anomalies from the data**

Anomaly detection using machine learning



Debreceni Egyetem Informatikai Kar

DL-supported anomaly detection

Training – masked reconstruction + Huber

- We set part of the input (mask_ratio) to zero, and the network only tries to reconstruct the omitted pixels.
- Loss function: Huber (Smooth L1) – robust against outliers.
- Early stopping: if the validation loss does not improve, we stop training.

Loss function (Huber / Smooth L1):

$$L_\delta(r) = \begin{cases} \frac{1}{2}r^2, & \text{if } |r| \le \delta \\ \delta(|r| - \frac{1}{2}\delta), & \text{otherwise} \end{cases}$$

$$r = X - \hat{X}$$

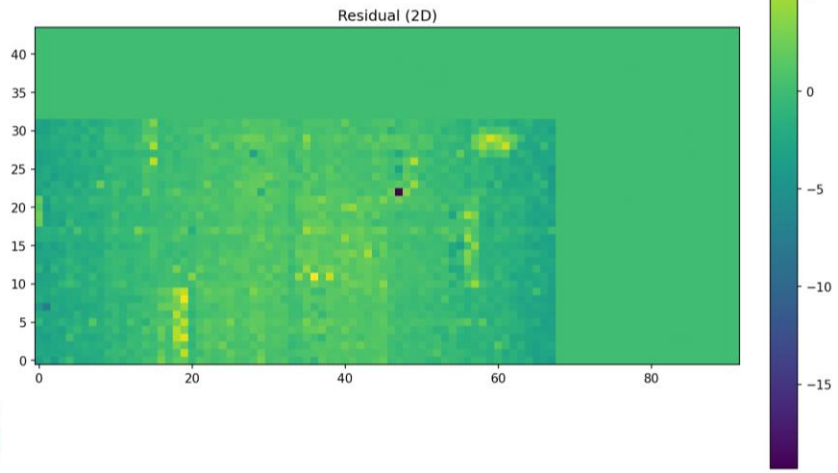
Anomaly detection supported by DL

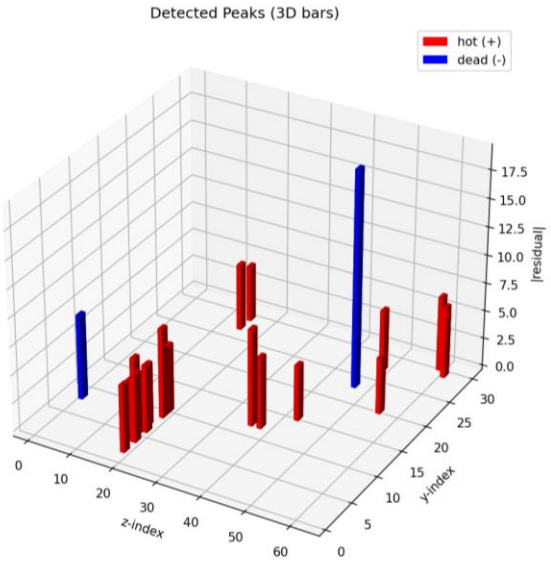
Conclusion – residual and threshold

- The learned model performs the reconstruction, the residu $R = X - \hat{X}$

- Based on the distribution of the residual, anomalies are marked with percentage thresholds:
  - hot: above the upper percentile of the positive range
  - dead: below the upper percentile of the negative range
  - extra hot: extreme cases

- Top 1% of largest errors is the recognized peak.





Debrecen

Thank you for your attention!