Machine Learning methods for NEWSdm data

Artem Golovatiuk
University of Kyiv (Ukraine)

Giovanni De Lellis,
Universita di Napoli (Italy),

Andrey Ustyuzhanin
Yandex (Russia)

Why do we need ML?

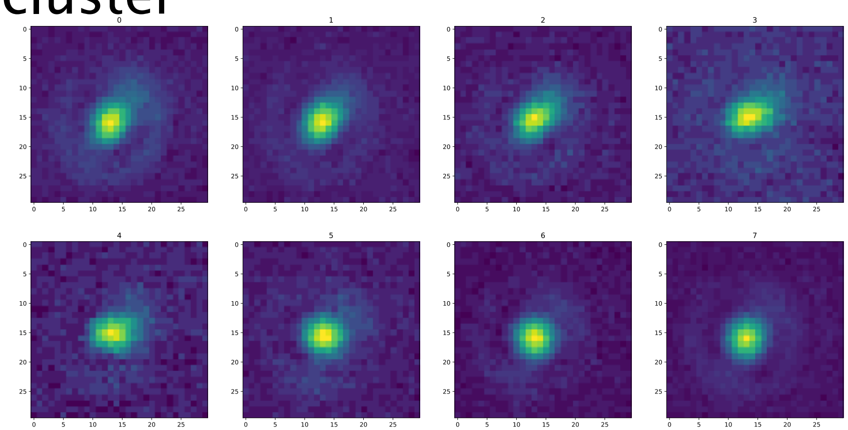
- Our goal:
  - Reducing the number of background events in potentially signal data
- Statistical approach:
  - limited by our physical understanding of the system
- Machine Learning approach:
  - can discover complex correlations between features, can be robust to insignificant variations in case of high input dimensions.

Algorithms performance metrics

- Algorithm's output – $Probability$.
- Common metrics in ML: ROC-AUC score
  Physically motivated metrics: Precision, Recall
- $$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$
- $$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$
- Use $Precision$ and $Recall$ to check the performance of the final algorithm
  - Need to select the probability threshold for the output.

Training data

- C 100keV signal ~ 15000 tracks
- LNGS exposed background ~ 7000 tracks
- Gaussian fit parameters (8 polarizations):
  - x, y – cluster center coordinates
  - l_x, l_y – major and minor axes of an elliptical fit
  - φ – direction of the cluster
  - n_{px} – area of the ellipse in pixels
  - br – brightness of the cluster
    - 56 features in total
- Cluster images:
  - 8 polarizations for each sample

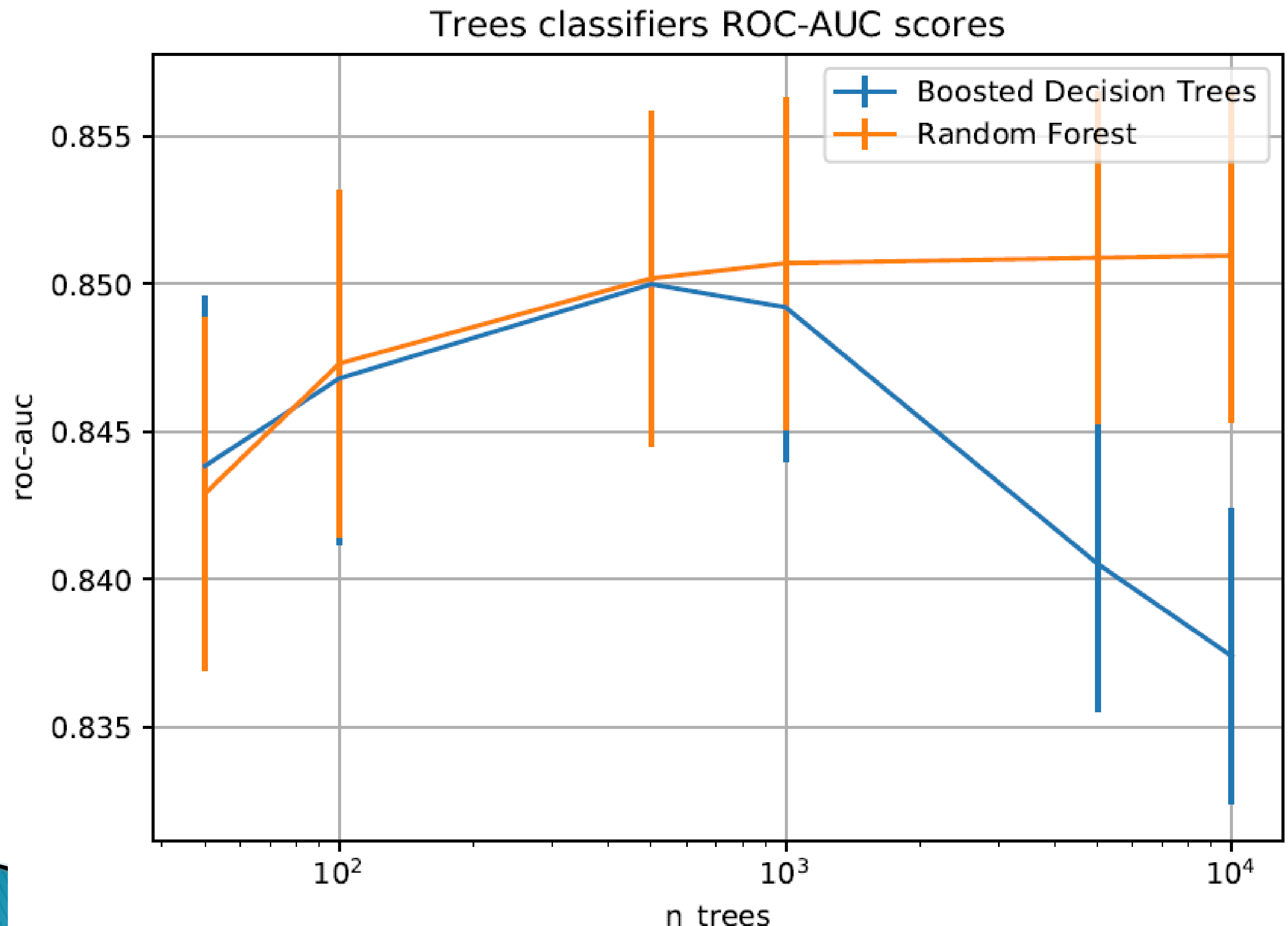


Tested approaches

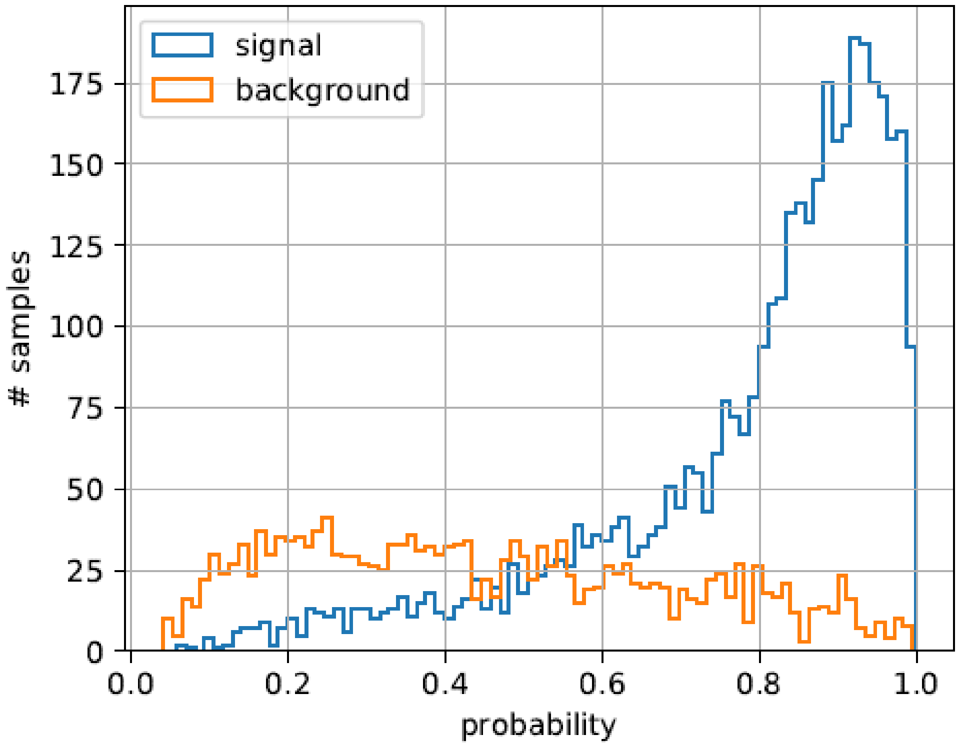
- Boosted Decision Trees:
  - Composition of small decision trees, next one improves result of the previous one.
  - Limited possibility to parallelize
- Random Forest:
  - Composition of very deep trees, each one makes its own decision, result is the average of probabilities.
  - Highly parallelizable on CPUs

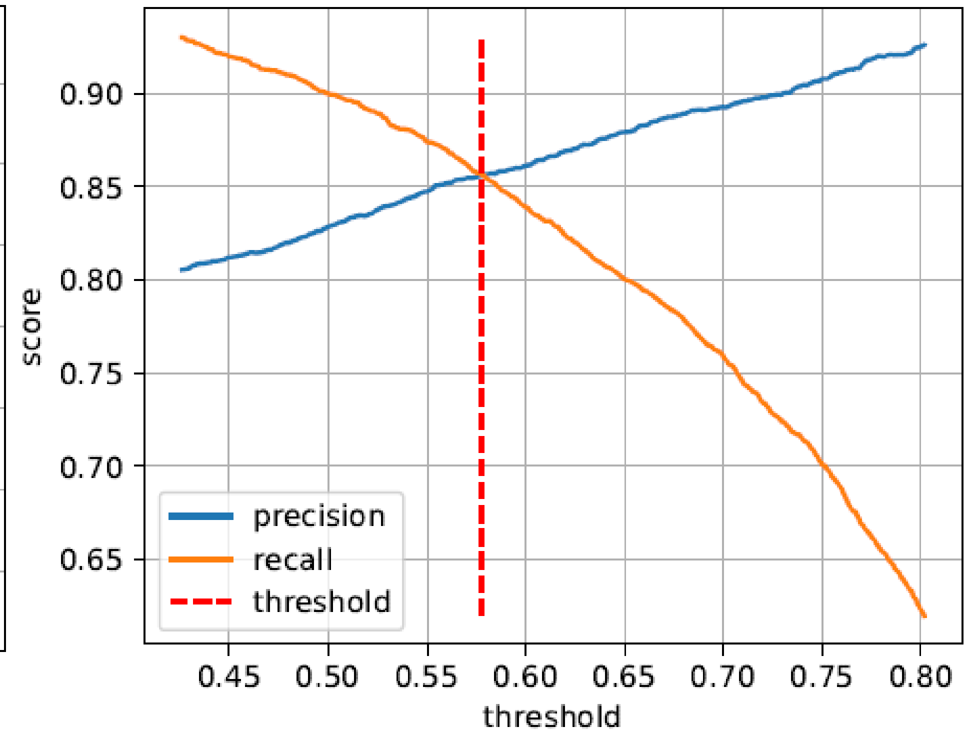
- Trees weakness:
  - Performance strongly depends on the features choice.

Preliminary results (Trees)



Preliminary results (Trees)



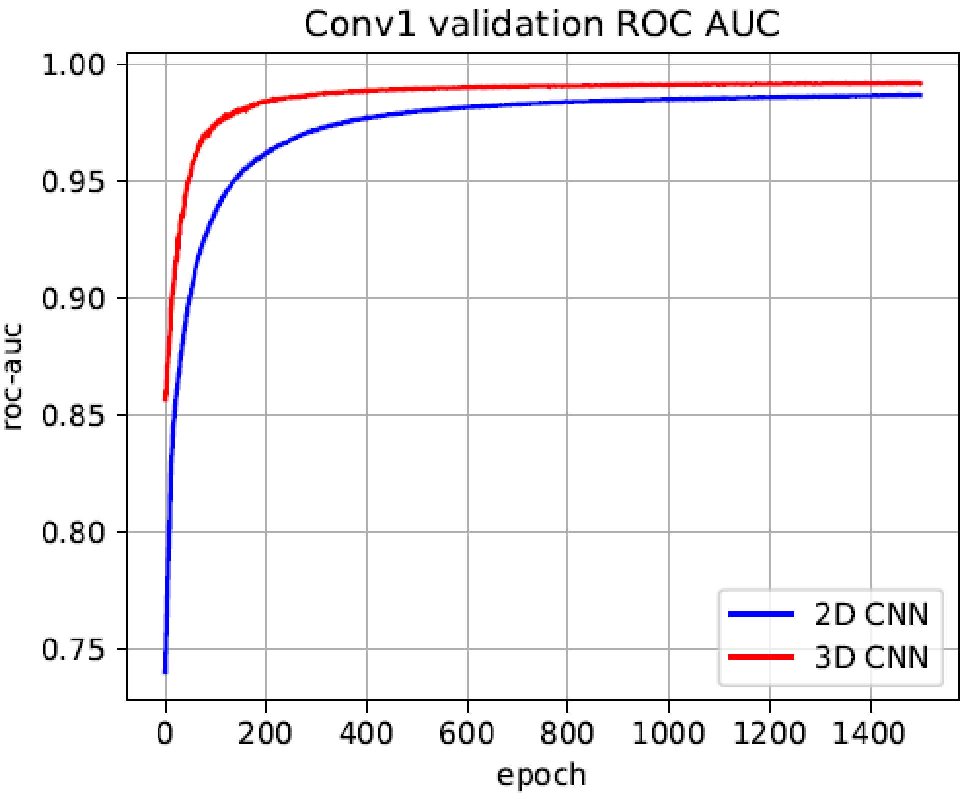


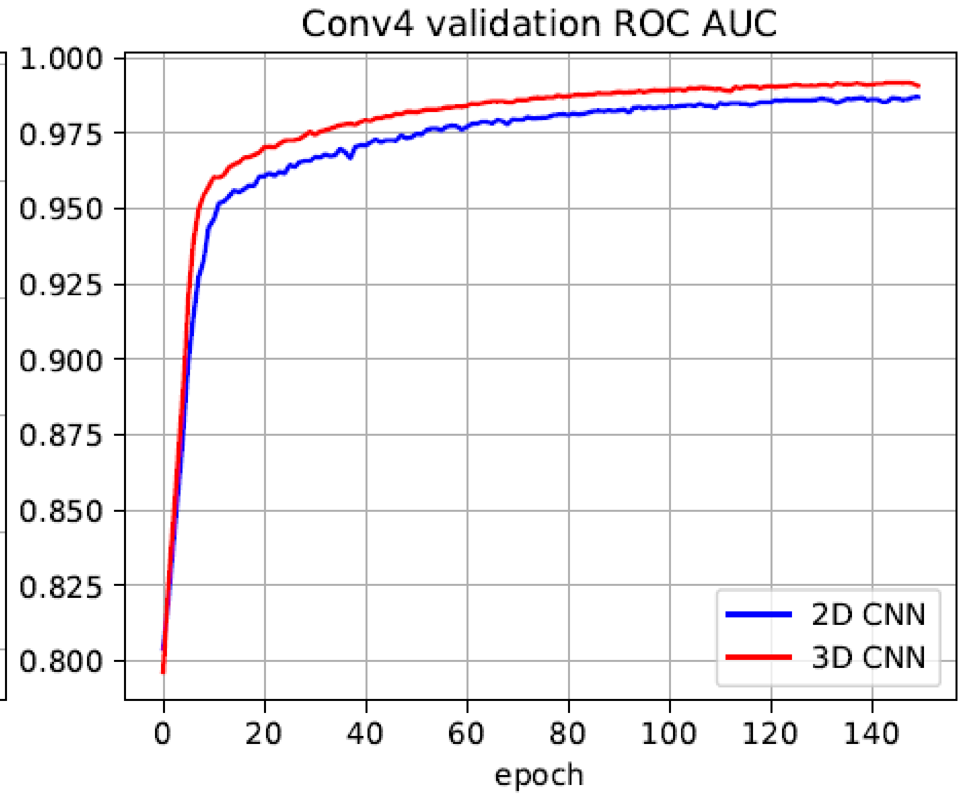
- Random Forest with 10^4 trees test output and physical scores

Tested approaches

- Convolutional Neural Networks:
    - Compared 2D and 3D architectures
    - Compared Deep and Shallow Networks
  - Working directly with the cluster images
  - Requires large computational power (e.g. GPU)
  - Larger datasets can be highly profitable for performance

Preliminary results (CNNs)

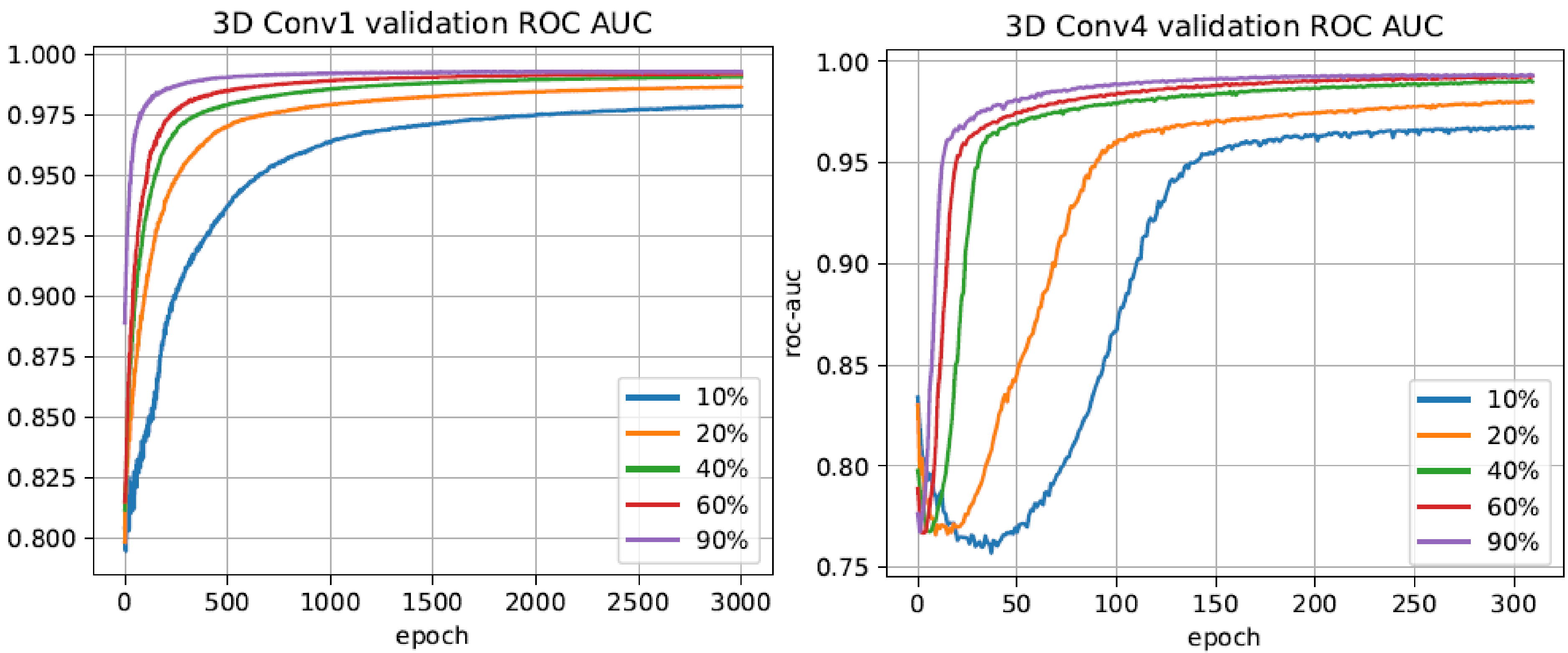




- Performance of Conv1 and Conv4 Networks

(named by the number of convolutional layers)

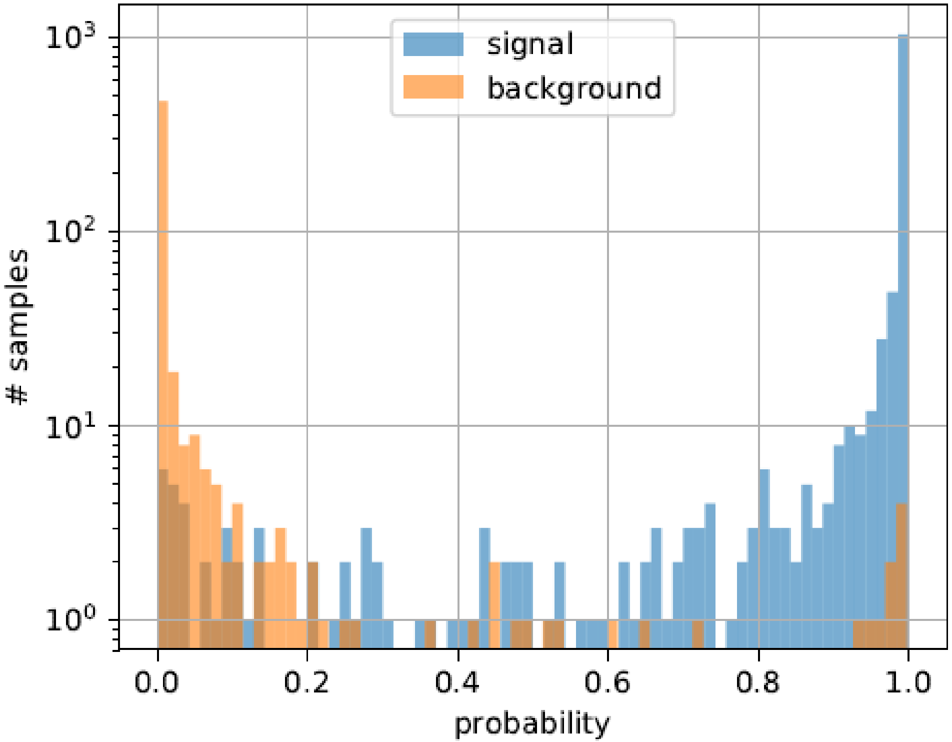
Preliminary results (CNNs)

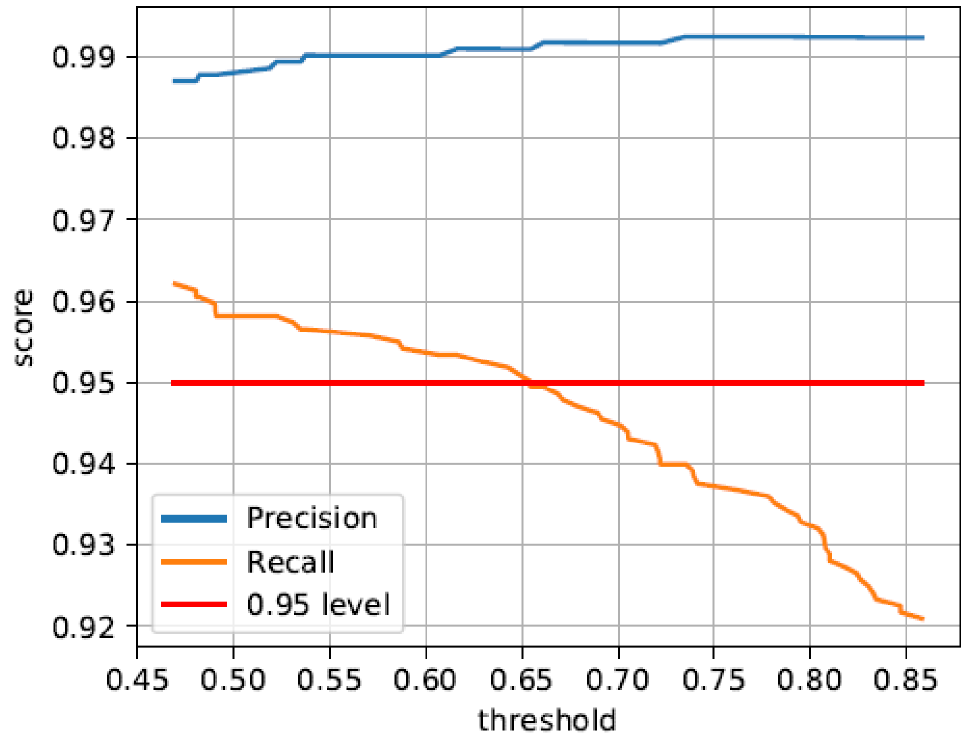


- Training size dependence of 3D Conv1 and Conv4 Networks

(named by the number of convolutional layers)

Preliminary results (CNNs)





- 3D Conv4 validation output and physical scores

Physical results

Score	Random Forest	3D Conv4 network
Precision	85.62%	99.01%
Recall	85.62%	95.41%

- Performance of the best algorithm in each class on the **test set** using the optimal threshold.
  - *Precision* lowers by background contamination.
  - *Recall* lowers by loosing the signal samples.

Conclusions and further plans

- Machine Learning can decrees background contamination by orders of magnitude without loosing any significant portion of signal.
- The physical motivation in the selection of network architecture can give considerable improvement of it’s performance.
- Enlarging the training set improves the networks’ performance.
- Further plans:
  - Enlarge and diversify the dataset.
  - Rotate the emulsions during scanning for isotropic signal.
  - Try getting the direction of the track as a physical feature.
  - Use images from color camera.