ER/NR discrimination

Initial plans and preliminary results

Luan G. M. de Carvalho

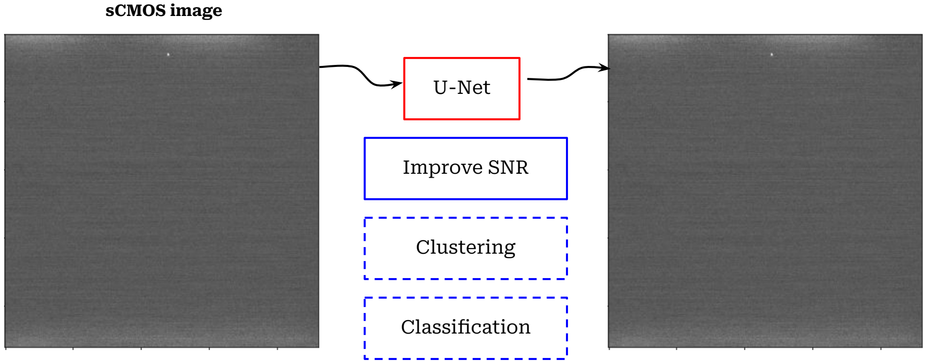
*with **Jordan Venâncio** (UFJF) and **Rafael A. Nóbrega** (UFJF)*

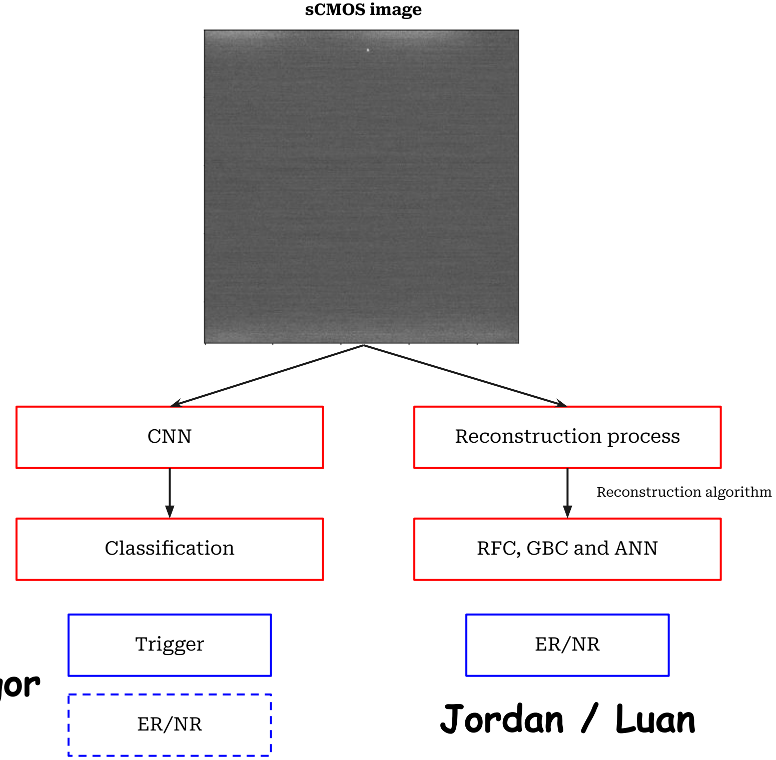
November 14, 2024

UFJF group

UFJF Machine Learning
→ on-going works

Guilherme





Igor

Jordan / Luan

Summary

Introduction

ER/NR discrimination

- Electron Recoil (ER) and Nuclear Recoil (NR) events are found in CYGNO Experiment.
- For **dark matter** searches, we are interested in NR events.

- Continue with Atul's work (A. A. Prajapati, "Multivariate Analysis for Background Rejection in CYGNO/INITIUM Experiment", PhD thesis (2024))
- Develop strategies to improve detection of such events
  - Signal efficiency (NR)
  - Background rejection (ER)

In CYGNO, each event is composed by:

- sCMOS image
- PMTs waveforms

Introduction

From Atul's thesis (A. A. Prajapati, "Multivariate Analysis for Background Rejection in CYGNO/INITIUM Experiment", PhD thesis (2024))

16 shape variables were used in machine learning algorithm training:

- energy
- size
- nhits
- length
- width
- slimness
- Gaussian Width
- LAPA
- thin track
- SDCD
- ChargeUnif
- MaxDen
- CylThick
- eta
- dE/dX
- dE/dA

Atul compared the discrimination performance for:

- Classical approach
- Deep Neural Networks
- Random Forest Classifier
- Gradient Boosting Classifier

Before pre-processing phase

Energies in a range from **2-50 keV**
25 energy levels
10000 events per level

Pre-processing phase (MC data)

- Geometrical Cut
- Noise Cut
- Outlier Removal from Integral Distribution

Atul's dataset

ER [keV]	NR [keV]	NR [keVee]	Events
2			10000
4	4	1.3	10000
6	6	2.5	10000
8	8	3.9	10000
10	10	5.4	10000
12	12	7	10000
14	14	8.7	10000
16	16	10.5	10000
18	18	12.2	10000
20	20	14	10000
22	22	15.9	10000
24	24	17.8	10000
26	26	19.6	10000
28	28	21.52	10000
30	30	23.42	10000
32	32	25.33	10000
34	34	27.25	10000
36	36	29.17	10000
38	38	31.1	10000
40	40	33	10000
42	42	34.98	10000
44	44	36.93	10000
46	46	38.88	10000
48	48	40.83	10000
50	50	42.80	10000
	100	92.23	1000
	200	191.9	1000
	300	291.8	1000
	400	391.75	1000
	500	491.7	1000
	600	591.7	1000
	700	691.67	1000
	800	791.66	1000
	900	891.65	1000
	1000	991.64	1000

Table 5.3: **Simulated** Energies and number of events for ER and NR.

Datasets

12 datasets - 1000 events each - 60 GB in total

Data = MC data no noise + Real pedestal

Energies:

ER [keV]	NR [keV]	Events
1	1	1000
3	3	1000
6	6	1000
10	10	1000
30	30	1000
60	60	1000

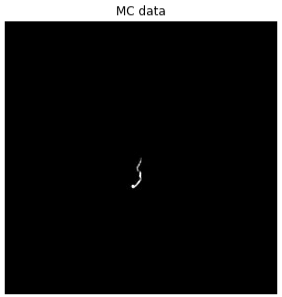
Total

6000 events of **ER**
6000 events of **NR**

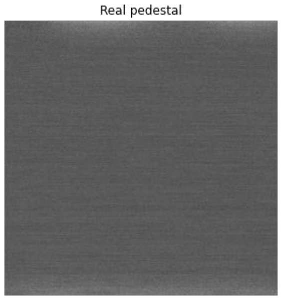
Not applying any
pre-processing to the data

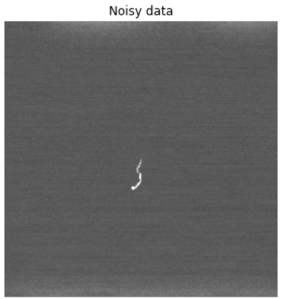
➔ Run reconstruction code to get variables from analysis

Example of event



+





Variables from reconstruction

Variable	Description
run	run number
event	event number
pedestal_run	run number used for pedestal subtraction
cmos_integral	integral counts of the full CMOS sensor
cmos_mean	average counts of the full CMOS sensor
cmos_rms	RMS of the counts of the full CMOS sensor
timestamp	Timestamp in UTC of the picture
t_DBSCAN	DBSCAN time
t_variables	Variables time
lp_len	# pixel
t_pedsub	pedestal subtraction
t_saturation	saturation correction mode
t_zerosup	zero suppression
t_xycut	xy acceptance cut
t_rebin	rebinning
t_medianfilter	median filter
t_noisered	noise reductor
nSc	nSc/i
sc_size	number of pixels of the cluster, without zero-suppression
sc_nhits	number of pixels of the cluster above zero-suppression threshold
sc_integral	uncalibrated integral of counts of all the pixels in the cluster
sc_corrintegral	density-corrected integral of the cluster (LEMON-specific calibration)
sc_rms	RMS of counts of all the pixels in the cluster
sc_energy	calibrated energy of the cluster in keV (LEMON-specific calibration)
sc_pathlength	curved length of the cluster (made with skeletonization)
sc_redpixIdx	index of the first pixel in the reduced pixel (redpix) collection belonging to the cluster
nRedpix	nRedpix/i
redpix_ix	x coordinate of the pixel
redpix_iy	y coordinate of the pixel
redpix_iz	number of counts of the pixel (after pedestal subtraction)
sc_theta	polar angle inclination of the major-axis of the cluster
sc_length	length of the major axis of the cluster
sc_width	length of the minor axis of the cluster
sc_longrms	truncated RMS of the cluster along the major axis

Variable	Description
sc_latrms	truncated RMS of the cluster along the minor axis
sc_lfullrms	full RMS of the cluster along the major axis
sc_tfullrms	full RMS of the cluster along the minor axis
sc_lp0amplitude	amplitude of the main peak of the longitudinal cluster profile
sc_lp0prominence	prominence of the main peak wrt the local baseline along the longitudinal cluster profile
sc_lp0fwhm	full width at half-maximum of the main peak of the longitudinal cluster profile
sc_lp0mean	mean position wrt the start of the cluster of the main peak of the longitudinal cluster profile
sc_tp0fwhm	full width at half-maximum of the main peak of the transverse cluster profile
sc_xmean	x position of the cluster energy baricenter
sc_ymean	y position of the cluster energy baricenter
sc_xmax	x position of the rightmost pixel of the cluster
sc_xmin	x position of the leftmost pixel of the cluster
sc_ymax	y position of the topmost pixel of the cluster
sc_ymin	y position of the bottommost pixel of the cluster
sc_pearson	Pearson coefficient of the cluster
sc_tgaussamp	amplitude of the Gaussian transverse profile
sc_tgaussmean	mean position of the Gaussian transverse profile
sc_tgausssigma	standard deviation of the Gaussian transverse profile
sc_tchi2	chi-squared of the Gaussian fit to the transverse profile
sc_tstatus	status of the Gaussian fit to the transverse profile
sc_lgaussamp	amplitude of the Gaussian longitudinal profile
sc_lgaussmean	mean position of the Gaussian longitudinal profile
sc_lgausssigma	standard deviation of the Gaussian longitudinal profile
sc_lchi2	chi-squared of the Gaussian fit to the longitudinal profile
sc_lstatus	status of the Gaussian fit to the longitudinal profile
Lime_pressure	Lime pressure
Atm_pressure	Atmosperic pressure
Lime_temperature	Lime temperature
Atm_temperature	Atmosheric temperature
Humidity	
Mixture_Density	

- 65 variables in total
- **33** variables selected at first (highlighted)
  - Features for the ML algorithms

Tasks

On going

- Evaluate performance using Random Forest and Gradient Boosting Classifiers -> **Luan**
- Evaluate performance using Deep Neural Networks -> **Jordan**
- Compare obtained results with Atul's thesis

Future Plans

- Fine tuning and improvement of the models

Training data with RFC and GBC

Hyperparameters

Random Forest Classifier

sklearn.ensemble.RandomForestClassifier()

'n_estimators' (default = 100) **number of trees in the forest**

'max_depth' (default = None)

'max_leaf_nodes' (default = None)

'max_features' (default = ‘sqrt’) **number of features to consider when looking for the best split**

Gradient Boosting Classifier

sklearn.ensemble.GradientBoostingClassifier()

'n_estimators' (default = 100) **number of boosting stages**

'max_depth' (default = 3)

'max_leaf_nodes' (default = None)

'max_features' (default = None)

'learning_rate' (default = 0.1) **shrinks the contribution of each tree**

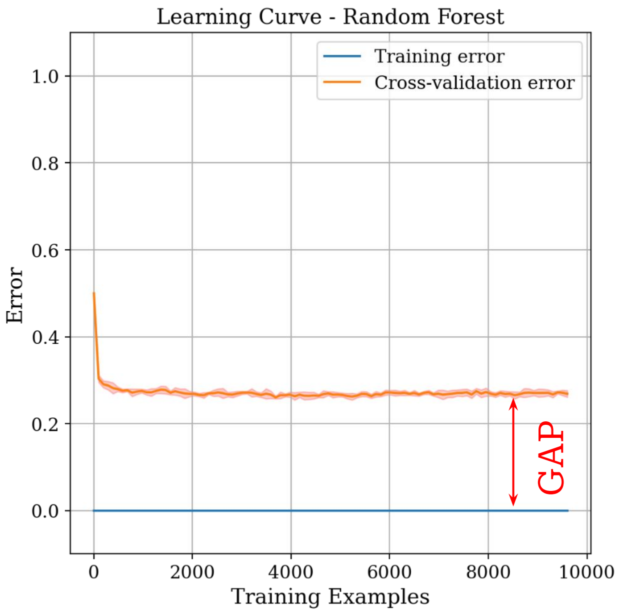
Training data with RFC and GBC

Learning curve using default hyperparameters

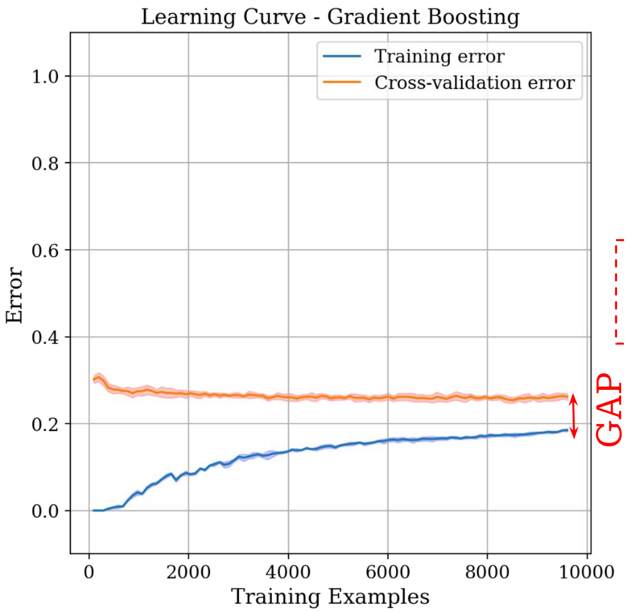
determines cross-validated training and test scores for different training set sizes

Will more data help performance get better?

Random Forest Classifier



Gradient Boosting Classifier



Search the best hyperparameters

Training data with RFC and GBC

HalvingGridSearchCV for best hyperparameters

sklearn.model_selection.HalvingGridSearchCV

Random Forest Classifier

sklearn.ensemble.RandomForestClassifier()

Grid to search

'n_estimators': range(10, 110, step = 10)

'max_depth': range(3, 9, step = 1)

'max_leaf_nodes': range(4, 50, step = 2)

'max_features': ['sqrt', None]

Best hyperparameters

'n_estimators': 30

'max_depth': 6

'max_leaf_nodes': 16

'max_features': 'sqrt'

Proceed the study with these optimal hyperparameters

Gradient Boosting Classifier

sklearn.ensemble.GradientBoostingClassifier()

Grid to search

'n_estimators': range(10, 110, step = 10)

'max_depth': range(3, 9, step = 1)

'max_leaf_nodes': range(4, 50, step = 2)

'max_features': ['sqrt', None]

'learning_rate': [0.01, 0.05, 0.1]

Best hyperparameters

'n_estimators': 70

'max_depth': 3

'max_leaf_nodes': 8

'max_features': 'sqrt'

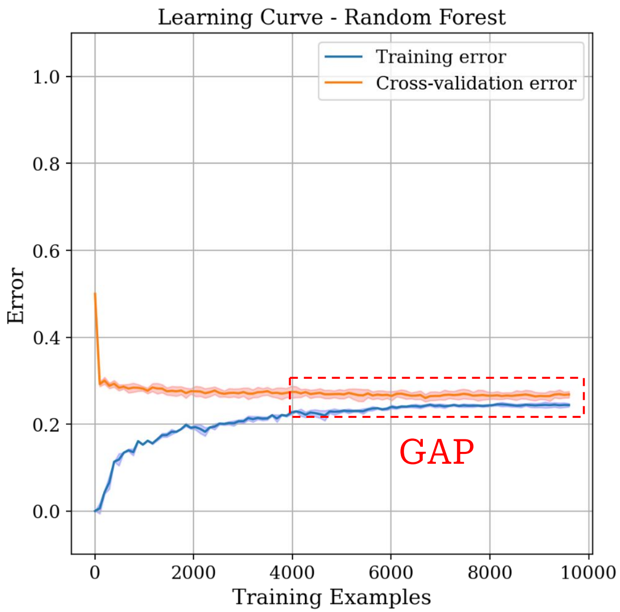
'learning_rate': 0.05

Training data with RFC and GBC

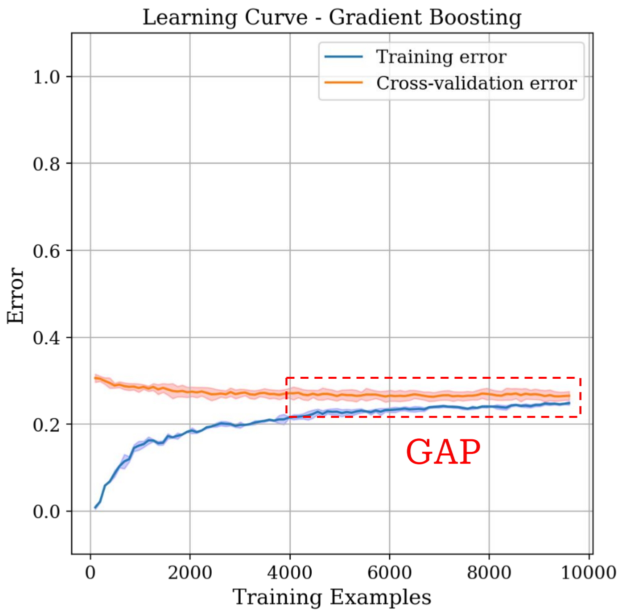
Learning curve using best hyperparameters

determines cross-validated training and test scores for different training set sizes

Random Forest Classifier

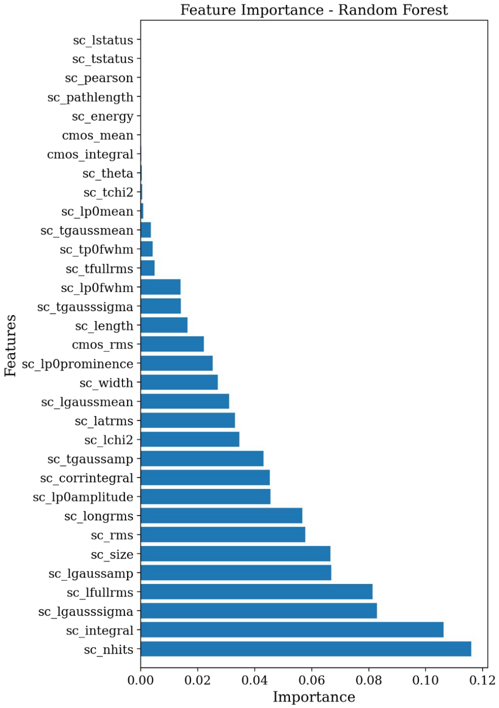


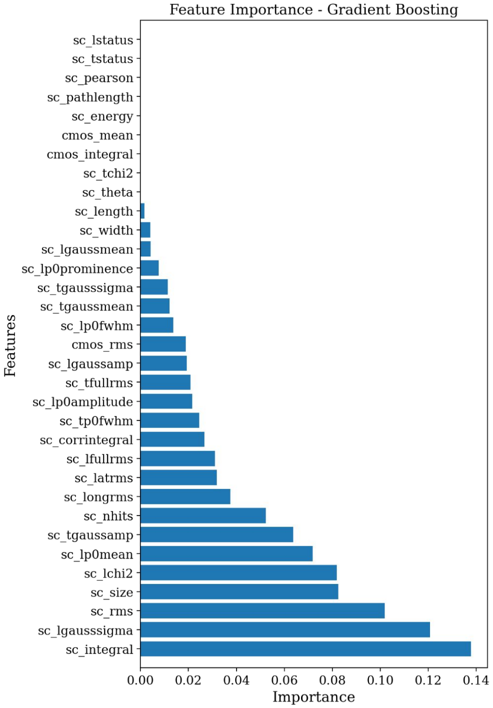
Gradient Boosting Classifier



Training data with RFC and GBC

Feature Importance using best hyperparameters





For this preliminary analysis

Let's remove the features with importance == 0

Training data with RFC and GBC

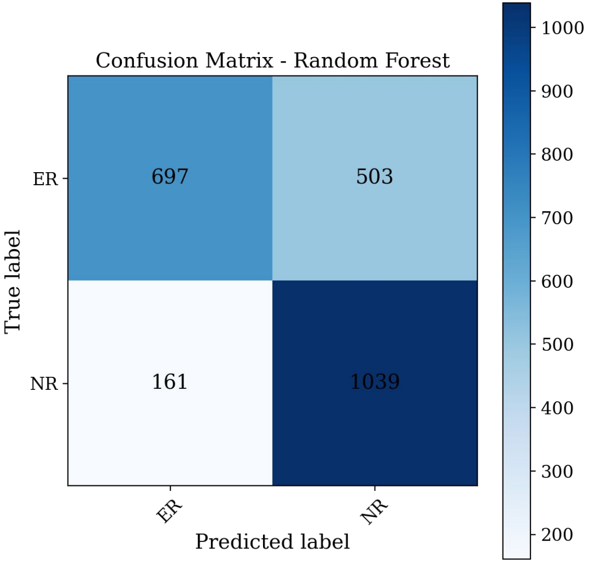
Variable	Description
run	run number
event	event number
pedestal_run	run number used for pedestal subtraction
cmos_integral	integral counts of the full CMOS sensor
cmos_mean	average counts of the full CMOS sensor
cmos_rms	RMS of the counts of the full CMOS sensor
timestamp	Timestamp in UTC of the picture
t_DBSCAN	DBSCAN time
t_variables	Variables time
lp_len	# pixel
t_pedsub	pedestal subtraction
t_saturation	saturation correction mode
t_zerosup	zero suppression
t_xycut	xy acceptance cut
t_rebin	rebinning
t_medianfilter	median filter
t_noisered	noise reductor
nSc	nSc/i
sc_size	number of pixels of the cluster, without zero-suppression
sc_nhits	number of pixels of the cluster above zero-suppression threshold
sc_integral	uncalibrated integral of counts of all the pixels in the cluster
sc_corrintegral	density-corrected integral of the cluster (LEMON-specific calibration)
sc_rms	RMS of counts of all the pixels in the cluster
sc_energy	calibrated energy of the cluster in keV (LEMON-specific calibration)
sc_pathlength	curved length of the cluster (made with skeletonization)
sc_redpixIdx	index of the first pixel in the reduced pixel (redpix) collection belonging to the cluster
nRedpix	nRedpix/i
redpix_ix	x coordinate of the pixel
redpix_iy	y coordinate of the pixel
redpix_iz	number of counts of the pixel (after pedestal subtraction)
sc_theta	polar angle inclination of the major-axis of the cluster
sc_length	length of the major axis of the cluster
sc_width	length of the minor axis of the cluster
sc_longrms	truncated RMS of the cluster along the major axis

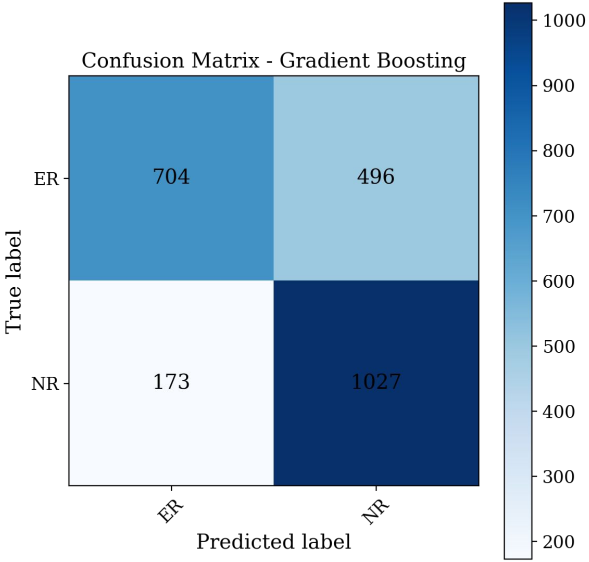
Variable	Description
sc_latrms	truncated RMS of the cluster along the minor axis
sc_lfullrms	full RMS of the cluster along the major axis
sc_tfullrms	full RMS of the cluster along the minor axis
sc_lp0amplitude	amplitude of the main peak of the longitudinal cluster profile
sc_lp0prominence	prominence of the main peak wrt the local baseline along the longitudinal cluster profile
sc_lp0fwhm	full width at half-maximum of the main peak of the longitudinal cluster profile
sc_lp0mean	mean position wrt the start of the cluster of the main peak of the longitudinal cluster profile
sc_tp0fwhm	full width at half-maximum of the main peak of the transverse cluster profile
sc_xmean	x position of the cluster energy baricenter
sc_ymean	y position of the cluster energy baricenter
sc_xmax	x position of the rightmost pixel of the cluster
sc_xmin	x position of the leftmost pixel of the cluster
sc_ymax	y position of the topmost pixel of the cluster
sc_ymin	y position of the bottommost pixel of the cluster
sc_pearson	Pearson coefficient of the cluster
sc_tgaussamp	amplitude of the Gaussian transverse profile
sc_tgaussmean	mean position of the Gaussian transverse profile
sc_tgausssigma	standard deviation of the Gaussian transverse profile
sc_tchi2	chi-squared of the Gaussian fit to the transverse profile
sc_tstatus	status of the Gaussian fit to the transverse profile
sc_lgaussamp	amplitude of the Gaussian longitudinal profile
sc_lgaussmean	mean position of the Gaussian longitudinal profile
sc_lgausssigma	standard deviation of the Gaussian longitudinal profile
sc_lchi2	chi-squared of the Gaussian fit to the longitudinal profile
sc_lstatus	status of the Gaussian fit to the longitudinal profile
Lime_pressure	Lime pressure
Atm_pressure	Atmosperic pressure
Lime_temperature	Lime temperature
Atm_temperature	Atmosheric temperature
Humidity	
Mixture_Density	

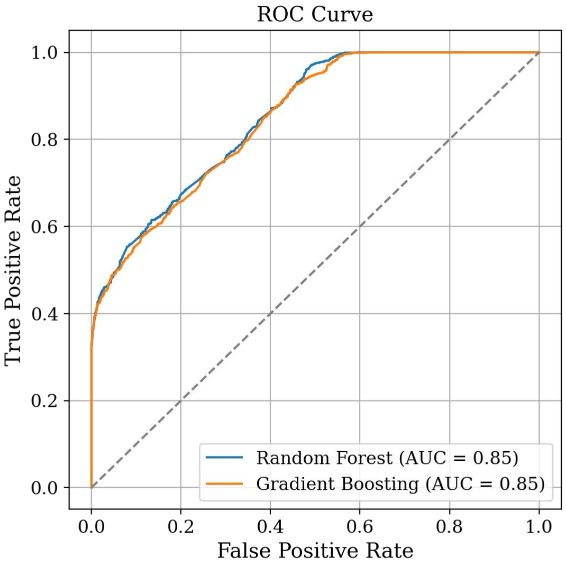
- **65** variables in total
- **33** variables selected at first (highlighted)
- **28** variables after selection based on the Feature Importance

Results with RFC and GBC

Training with all energies and test with all energies - 80% for training and 20% for test



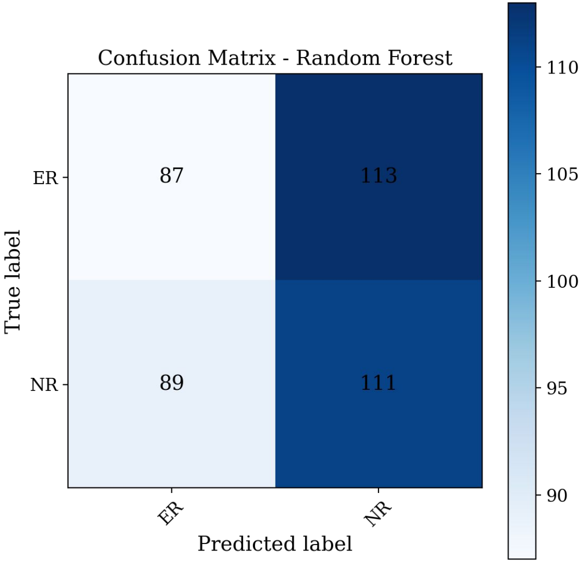


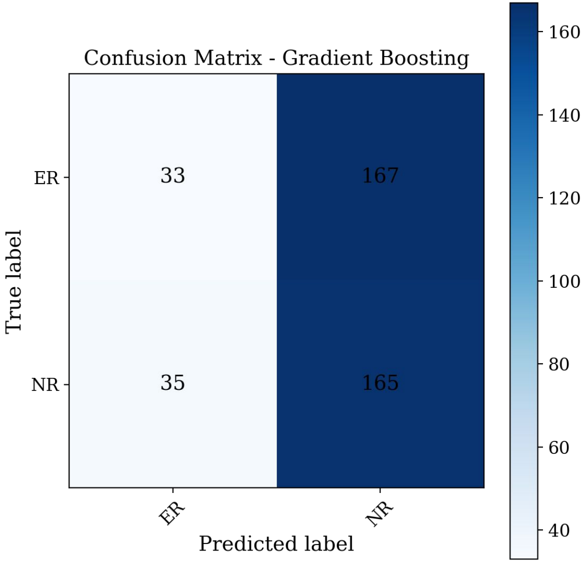


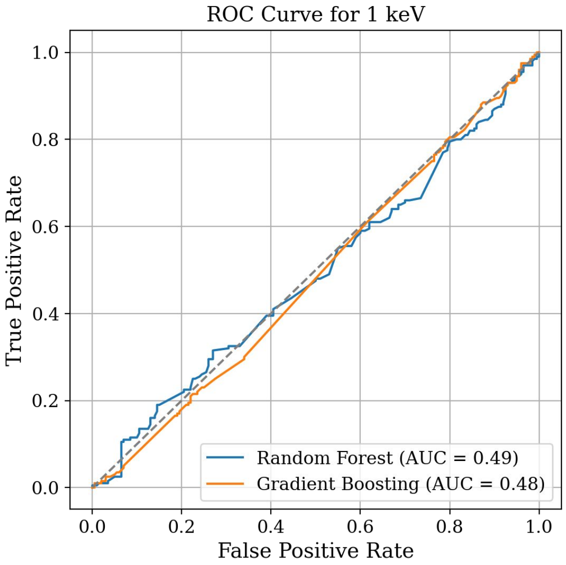
NR as positive class
ER as negative class

Results with RFC and GBC

Training with all energies and test with 1 keV - split 80% for training and 20% for test



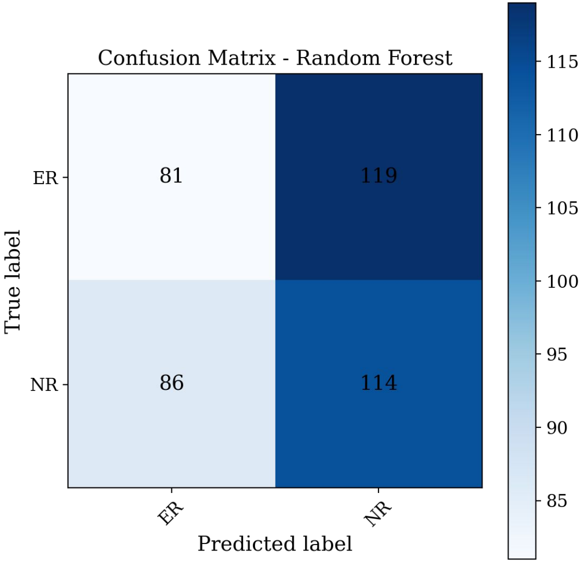


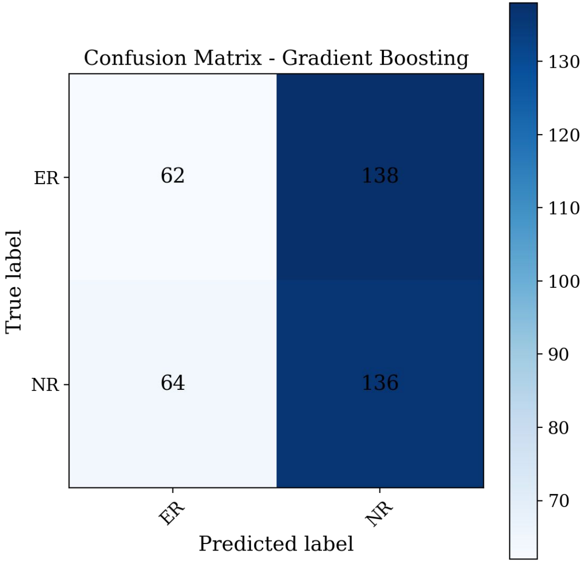


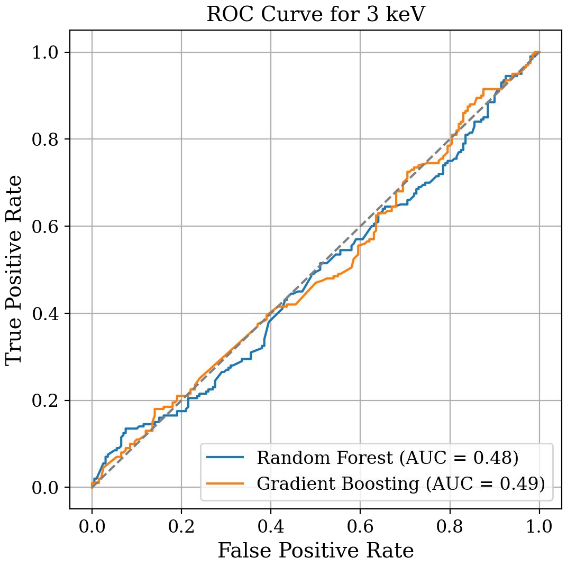
NR as positive class
ER as negative class

Results with RFC and GBC

Training with all energies and test with 3 keV - split 80% for training and 20% for test



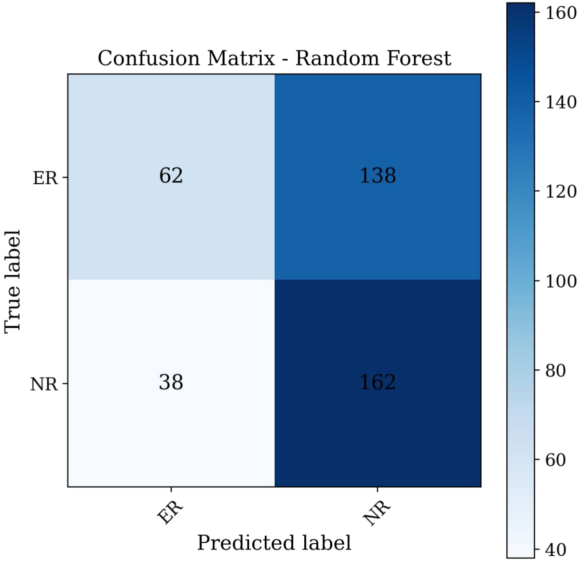


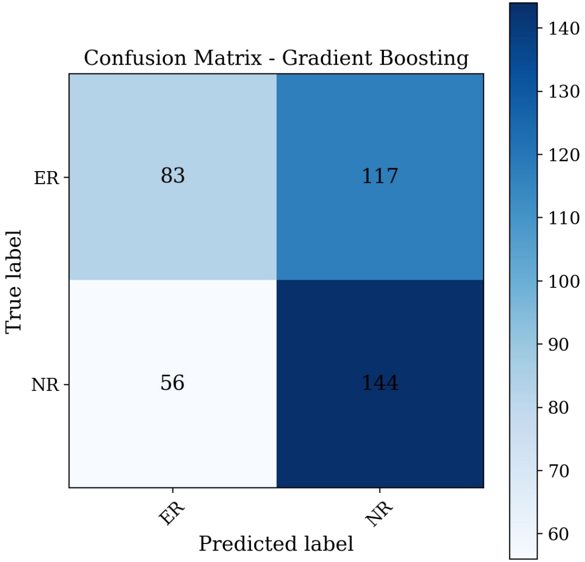


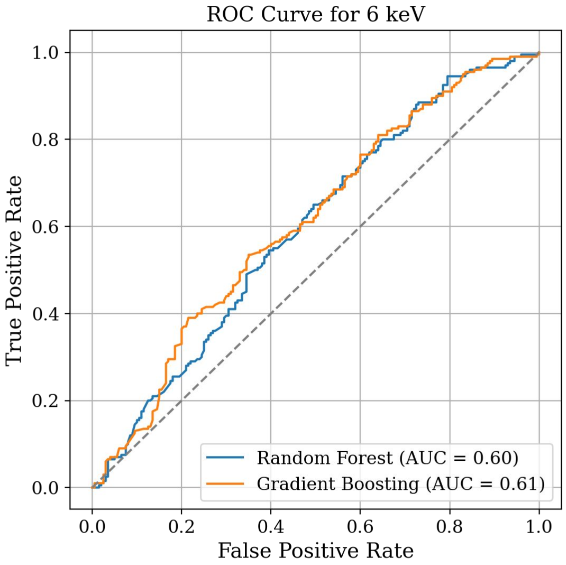
NR as positive class
ER as negative class

Results with RFC and GBC

Training with all energies and test with 6 keV - split 80% for training and 20% for test



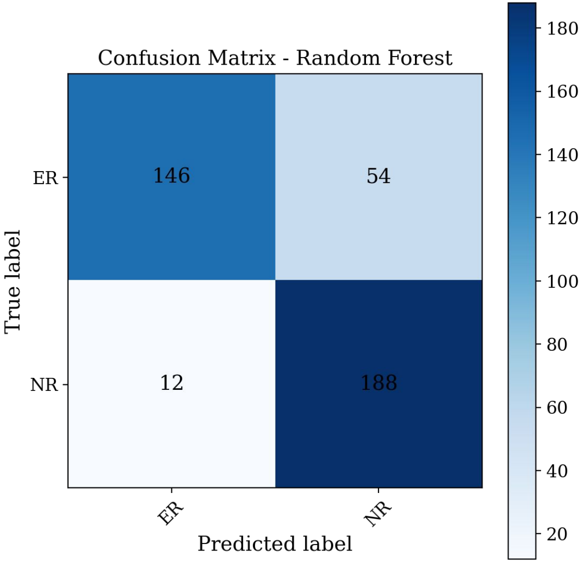


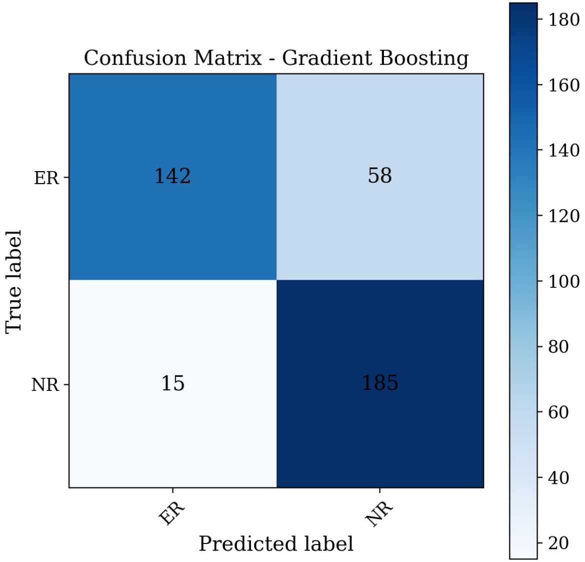


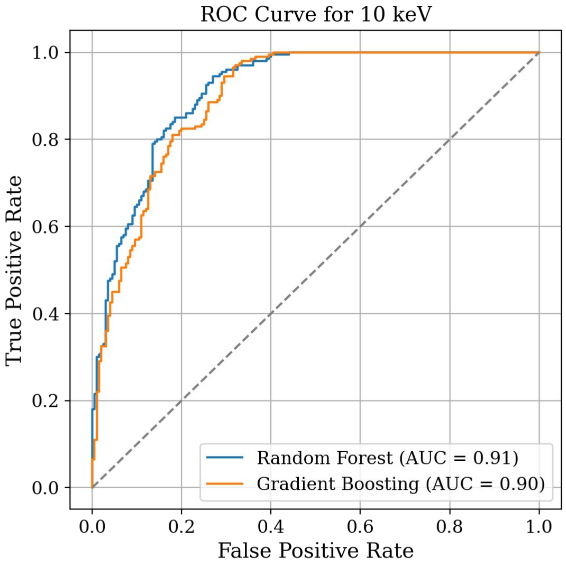
NR as positive class
ER as negative class

Results with RFC and GBC

Training with all energies and test with 10 keV - split 80% for training and 20% for test



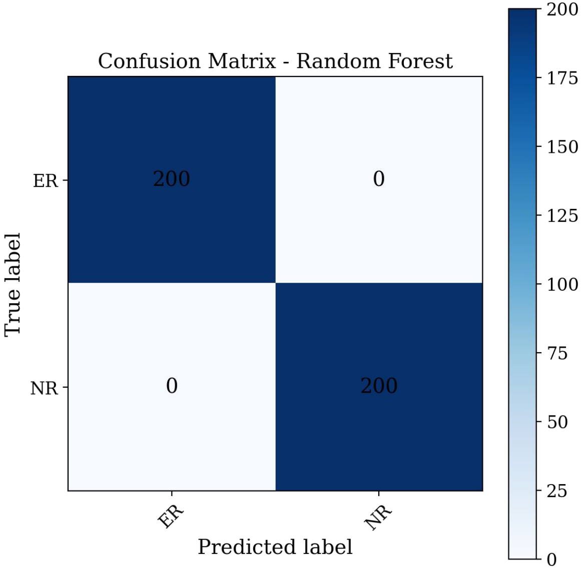


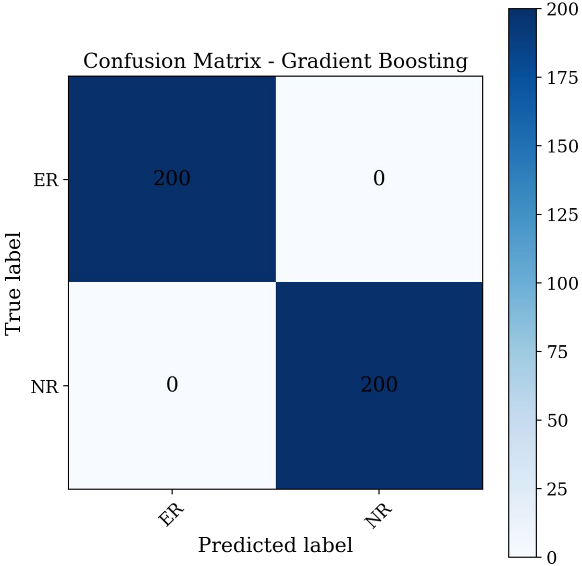


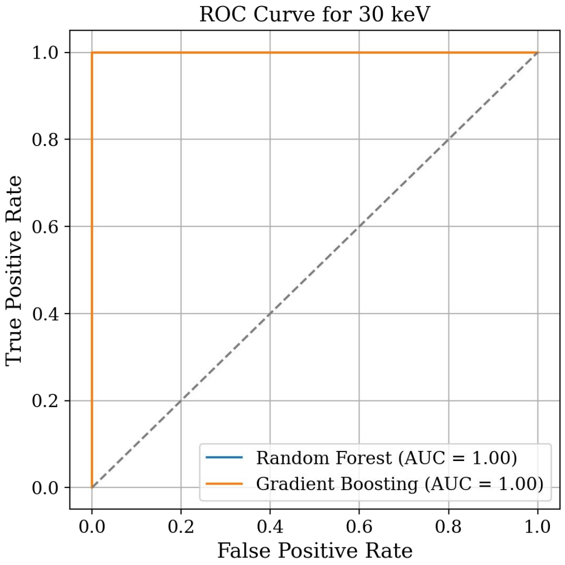
NR as positive class
ER as negative class

Results with RFC and GBC

Training with all energies and test with 30 keV - split 80% for training and 20% for test



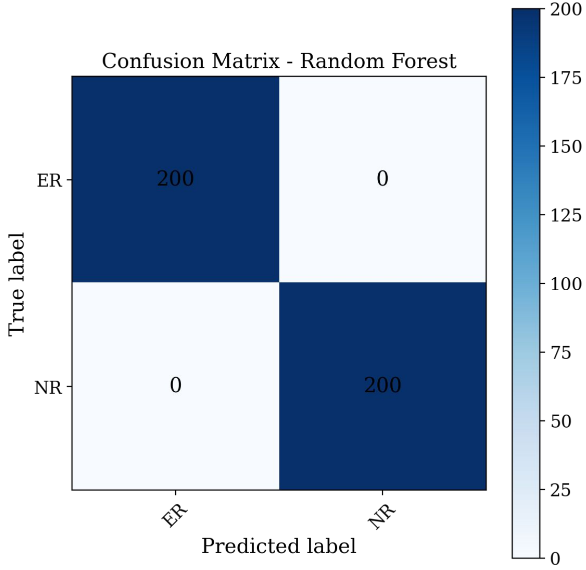


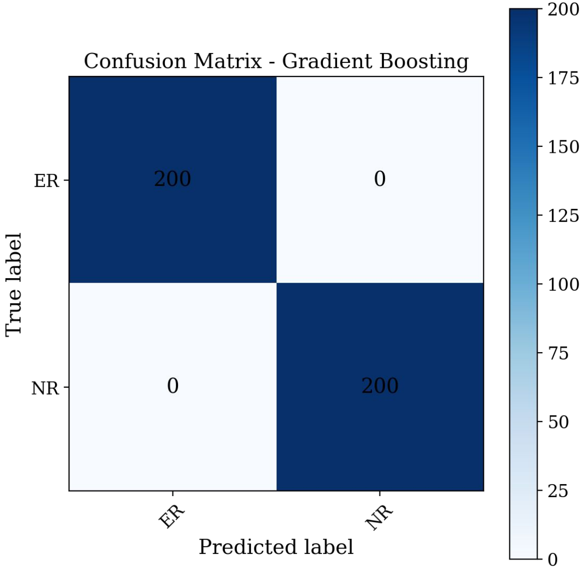


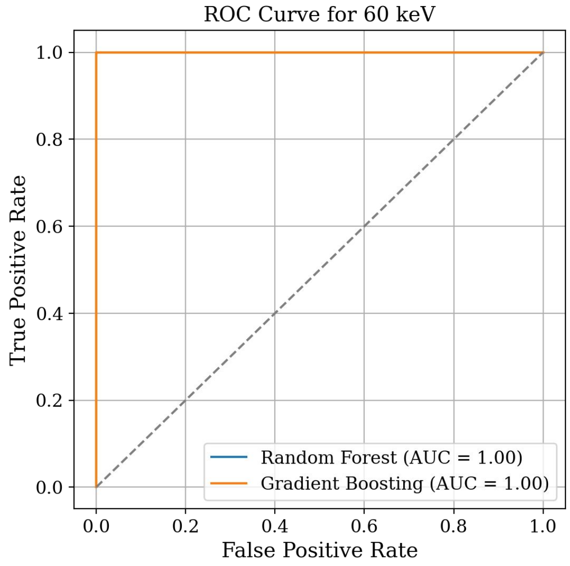
NR as positive class
ER as negative class

Results with RFC and GBC

Training with all energies and test with 60 keV - split 80% for training and 20% for test



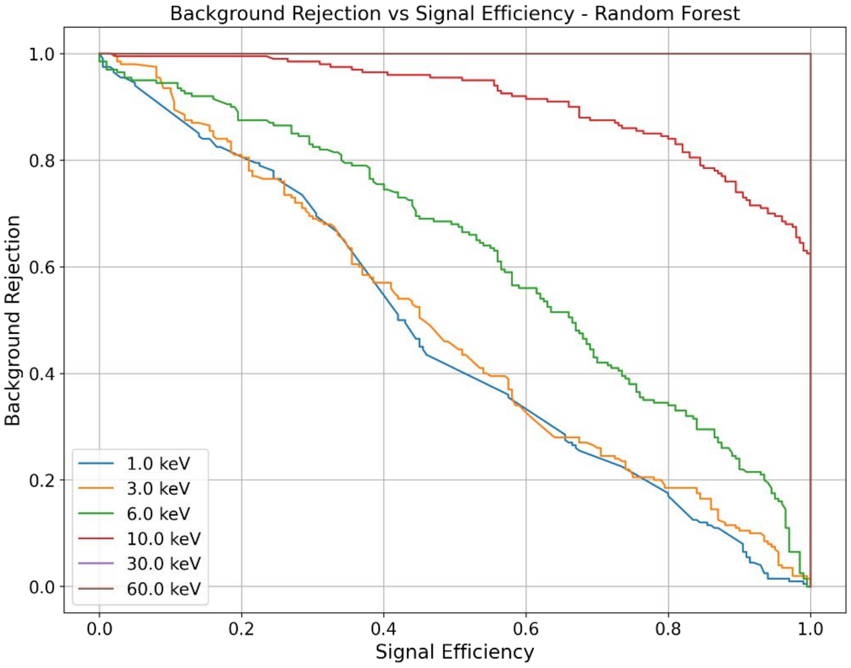


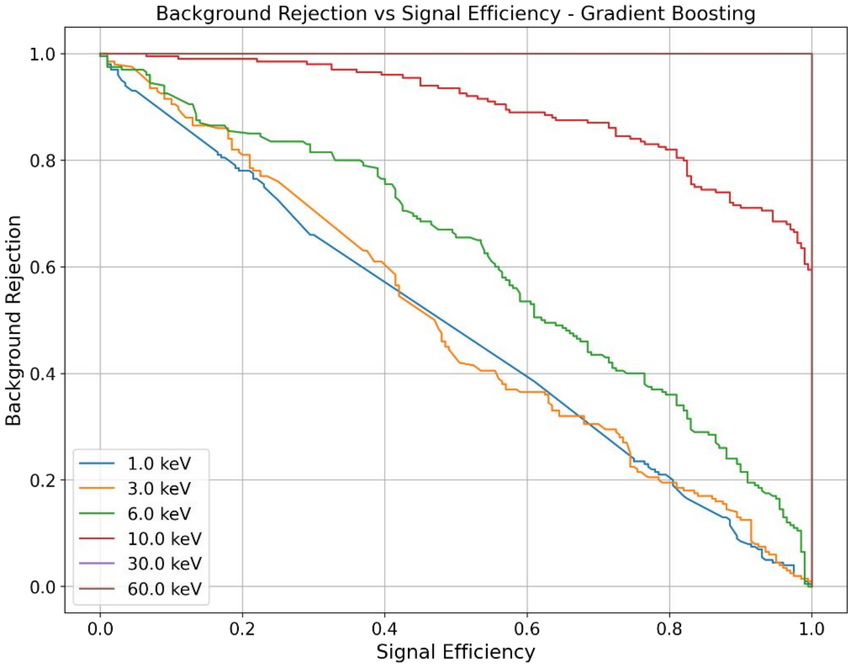


NR as positive class
ER as negative class

Results with RFC and GBC

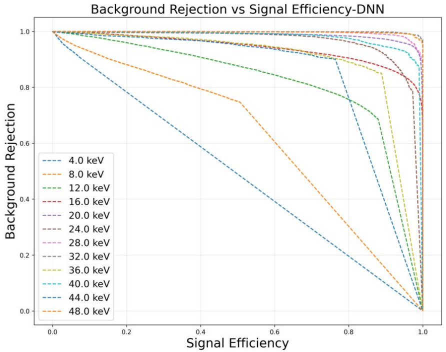
Training with all energies - Background Rejection vs Signal Efficiency

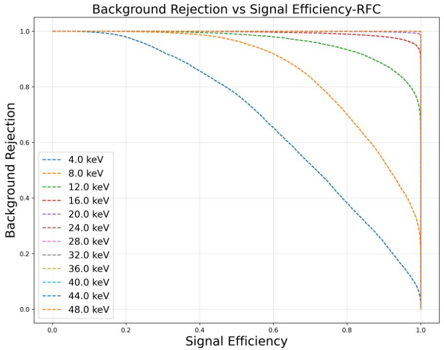


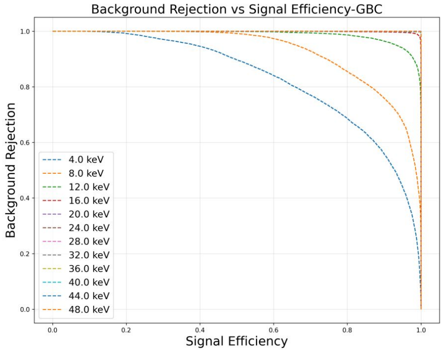


Results with RFC and GBC

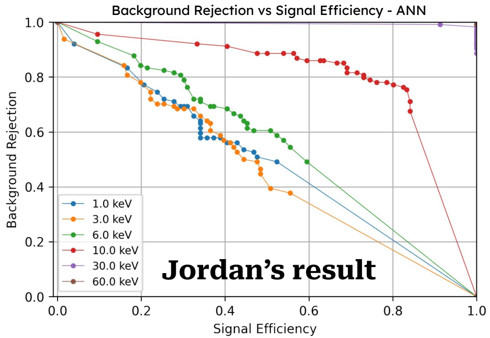
Training with all energies - Background Rejection vs Signal Efficiency

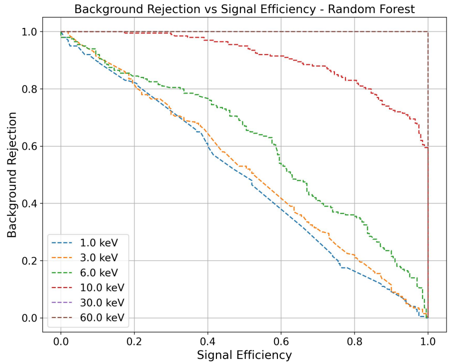


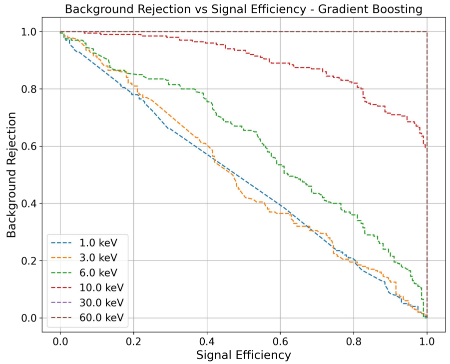




From Atul's thesis
Figure 6.19

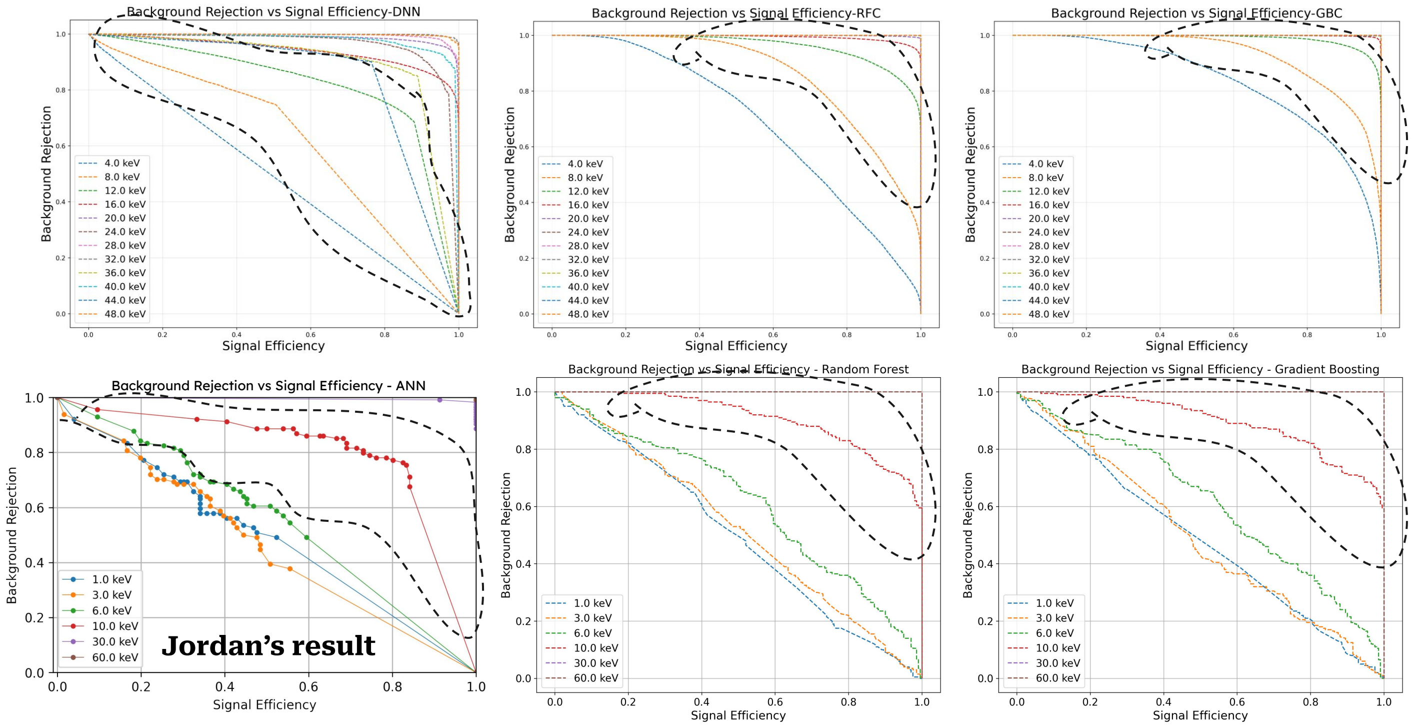






Results with RFC and GBC

Training with all energies - Background Rejection vs Signal Efficiency



From Atul's thesis
Figure 6.19

Comparing 10 keV

Conclusions

- A preliminary study has been done for ER/NR classification
  - RFC, GBC, ANN
- Results showed divergences and similarities with Atul's results
  - The comparison is not trivial and direct

Atul's dataset

Energies in a range from **2-50 keV**
25 energy levels
~10000 events per level

My dataset

Energies in a range from **1-60 keV**
6 energy levels
1000 events per level

Jordan's result

Jordan's **ANN** performed slightly better than Atul's **DNN**

- **Without applying:**
  - Pre-processing
  - Robust feature selection
  - Data augmentation

Next steps → Evaluate performance