EUROfusion

Advancing Interpretability in Artificial Intelligence Disruption Prediction Models: A Cross-Tokamak Perspective

E. Aymerich, B. Cannas, A. Fanni, F. Pisano, G. Sias *, the JET Contributors, the WPTE Team

(*) Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

IAEA-Third Technical meeting on Plasma Disruptions and their Mitigation
3rd-6th September, ITER Headquarters
Saint-Paul-lez-Durance, France

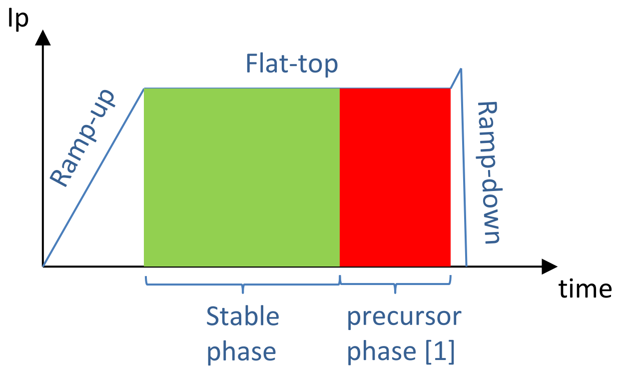
This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200 — EUROfusion). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Outline

- JET Dataset
- Disruption prediction by ML algorithms
  - ✓ CNN predictor
  - ✓ Comparison among CNN, MLP and GTM
  - ✓ CNN upgrade with vertical bolometer
- Unsupervised disruption predictor
  - ✓ SOM predictor
- Conclusions and ongoing work

JET Dataset - UNICA

Plasma parameters	Acronym	Dimensionality	Diagnostics
Electron Temperature *profile*	T_e	*1-D*	HRTS
Electron Density *profile*	n_e	*1-D*	HRTS
Radiated Power *profile*	P_{rad}	*1-D*	Bolometer (H, V)
Total Radiated Power	$P_{rad-TOT}$	*0-D*	Bolometer
Total Input Power	P_{TOT}	*0-D*	BetaLi
Internal Inductance	l_i	*0-D*	BetaLi
Normalized locked mode	LM_{norm}	*0-D*	LMS
MHD spectrogram	Spectr	*1-D*	Mirnov coils



[1] E. Aymerich *et al*, ***Nuclear Fusion 2021***, 61(3), 036013

Te, ne and Prad peaking factors (0D):

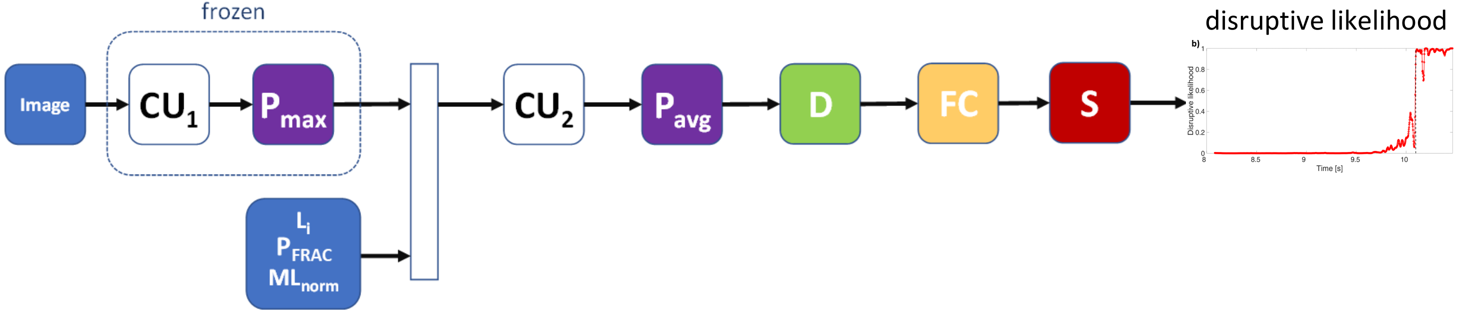
- encode spatial information
- defined heuristically
- lose information as they spatially average profile values

Te, ne and Prad profile images (1D):

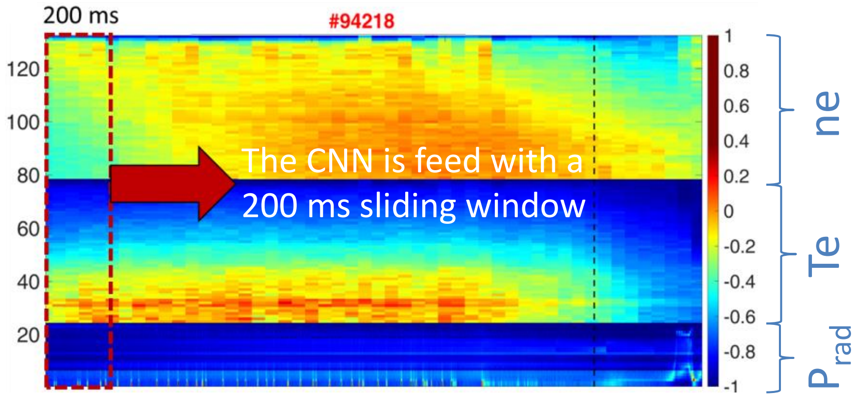
- no heuristic definition
- reports the entire profile values

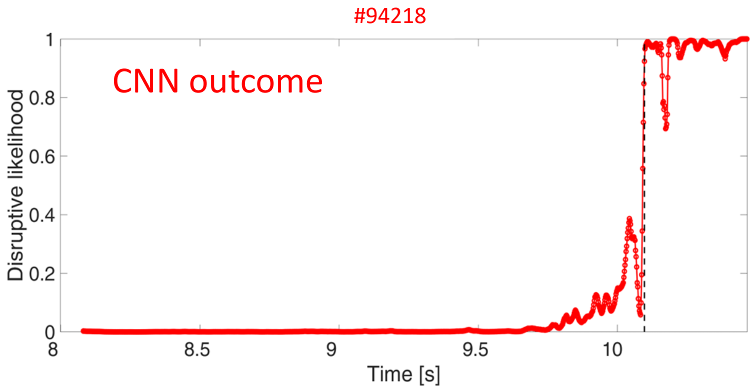
set	Campaigns		Disruptions	Regular
I	**2011÷2013**	C28-C30	**127**	**115**
II	**2016**	C36	**29**	**41**
III	**2019÷2020**	C38	**37**	**63**

CNN predictor architeture [2]



Te, ne and Prad profiles have been treated as a single image. l_i, P_{frac} and LM_{norm} are fed in downstream of the first filter block.





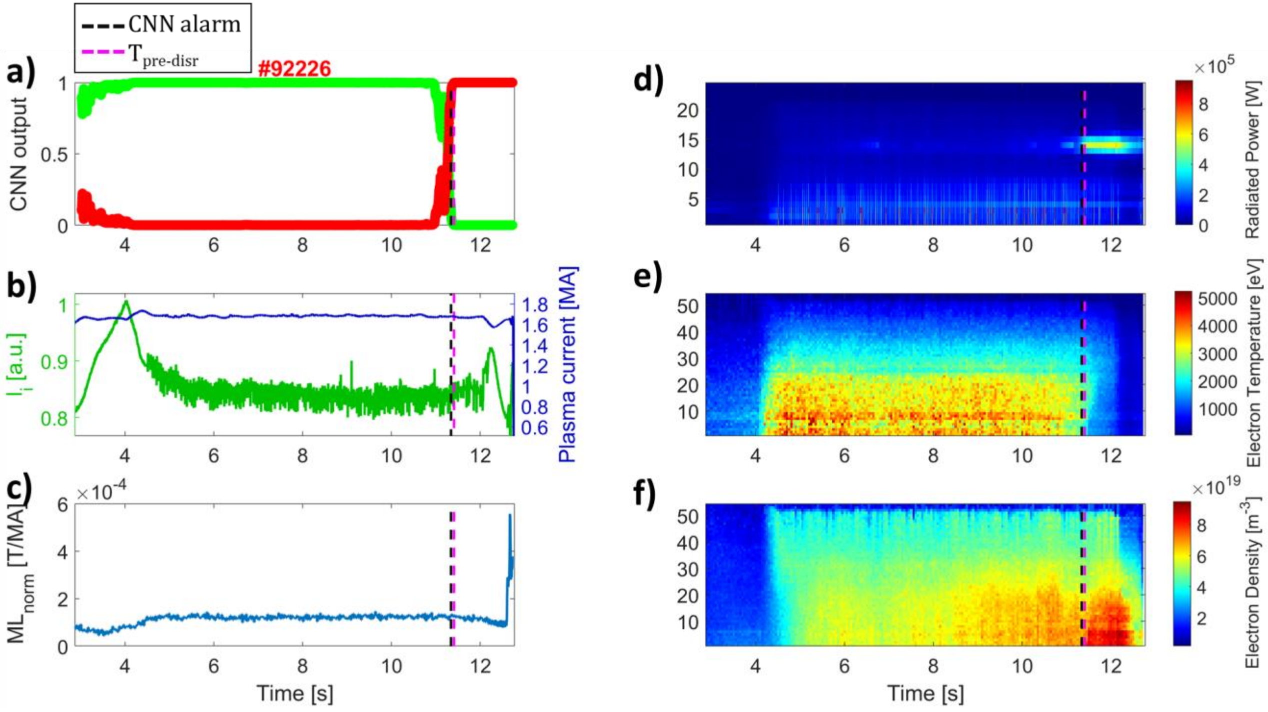
[2] E. Aymerich *et al.*, Nucl. Fusion, vol. 62, p. 066005, 2022.

CNN disruptive pattern [2]

Paths responsible for the alarms can be easily identified

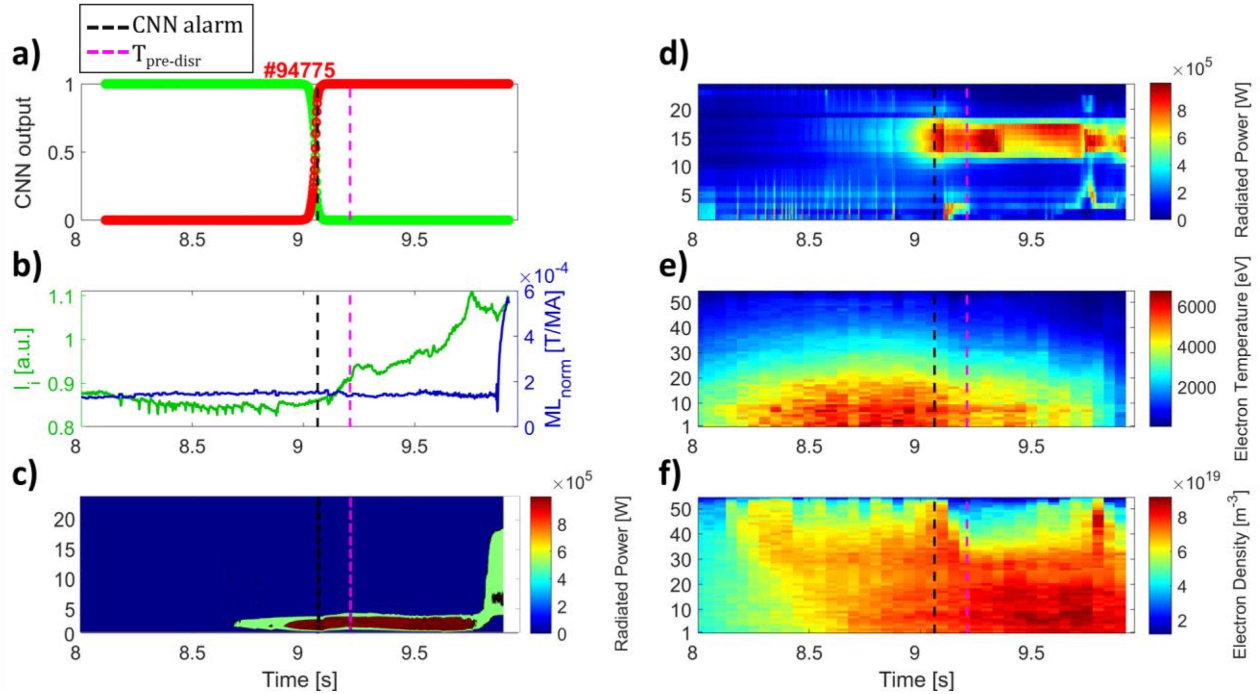
- ❑ Impurity accumulation disruptive mechanism:

- ✓ strong radiation from the central chords of BOL-H (Figure d)
- ✓ electron temperature collapse at plasma core (LOS <23, Figure e)
- ✓ core electron density peaking (LOS <22, Figure f).



CNN disruptive pattern [2]

Paths responsible for the alarms can be easily identified

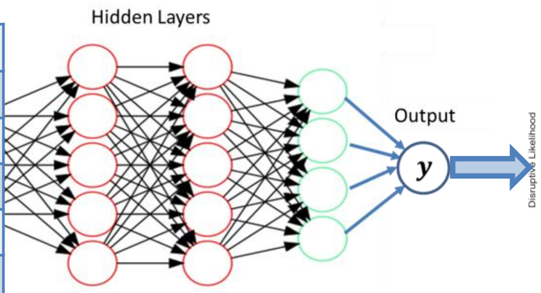


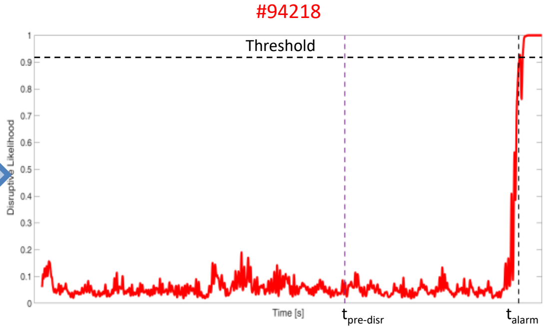
- Edge collapse disruptive mechanism:
  - ✓ rise of the plasma internal inductance (Figure b).
  - ✓ radiation at the central chords of BOL-H (Figure d)
  - ✓ cooling of the plasma between LOS 12 and 30 of HRTS (R between 3.13m to 3.46m, Figure e)

The further analysis of the BOL-V data allows to localize the radiation blob in the outboard of the plasma (chords 1-5, Figure c).

MLP and GTM predictor architetures [3]

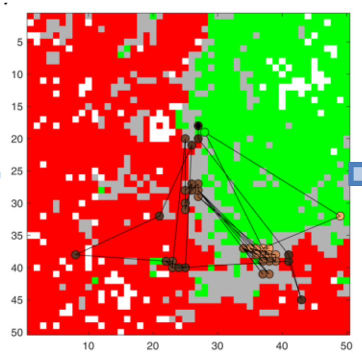
Peaking factor of temperature	Te_{pf}
Peaking factor of electron density	ne_{pf}
Peaking factor_1 of the radiation(*)	$RAD_{pf\text{-}CVA}$
Peaking factor_2 of the radiation(**)	$RAD_{pf\text{-}XDIV}$
Internal Inductance	li
Normalized locked mode	LM_{norm}

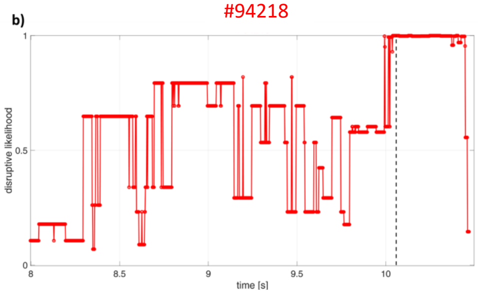


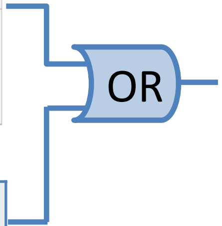


Multiple condition alarm scheme of the GTM predictor proposed [4]

Peaking factor of temperature	Te_{pf}
Peaking factor of electron density	ne_{pf}
Peaking factor_1 of the radiation(*)	$RAD_{pf\text{-}CVA}$
Peaking factor_2 of the radiation(**)	$RAD_{pf\text{-}XDIV}$
Internal Inductance	li







Normalized locked mode thresholding

The peaking factors are defined as a '*core versus all*' metric [4]
(*) excluding the X-point/divertor region from all
(**) excluding the core region from all

[4] A Pau et Al., Nucl. Fusion 2019, 59, 106017

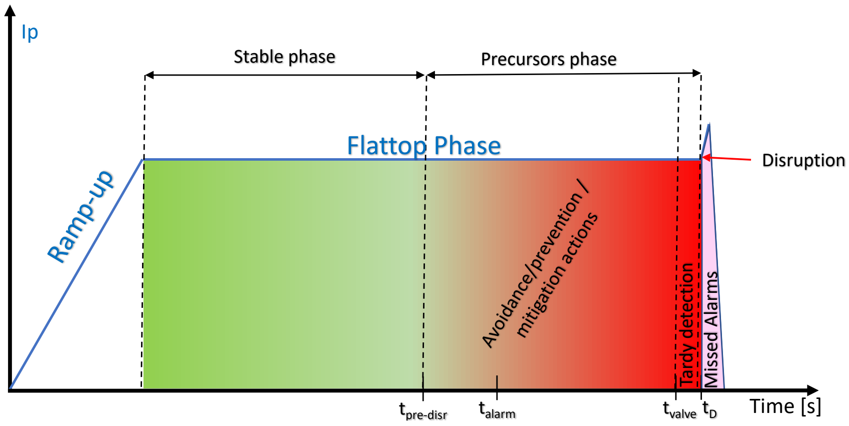
Predictor performance comparison [3]

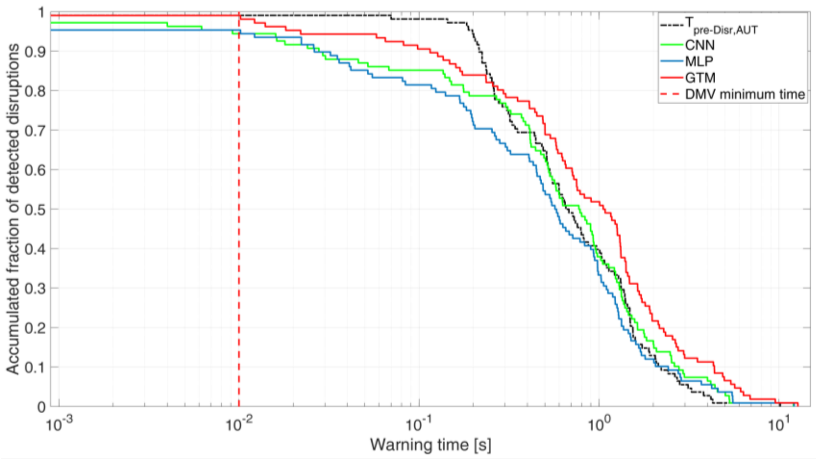
set	Campaigns	Disruptions	Regular
Training	C28-C30	85	70
Test	C36,C38	108	149

Performance index	MLP	GTM	CNN
SP-test[%]	95.37	97.22	94.44
MA-test [%]	2.78	1.85	2.78
FA-test [%]	3.36	18.79	5.37
Feature extraction	Manual	Manual	**Automatic**
Interpretability	Black box	**Interpretable**	Black box

- Successful prediction (SP), Missed Alarms (MAs), False alarms (FAs)
- Cumulative fraction of predicted disruptions:

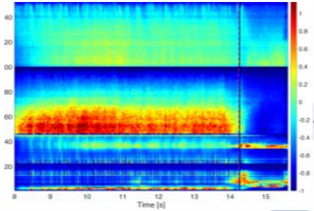
reports the value, in per unit, of successful alarms activated before the corresponding $warning\ time\ (tD - talarm)$

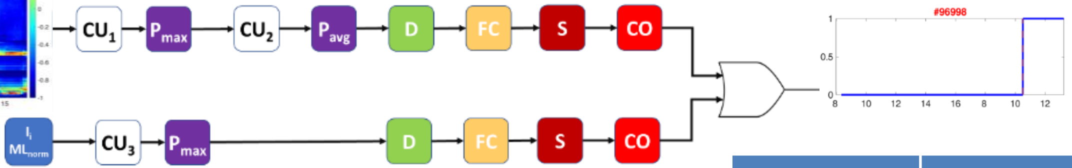




CNN predictor upgrade adding vertical bolometer data [5]

[5] E. Aymerich *et al.*, Fusion Engineering and Design 193 (2023) 113668.

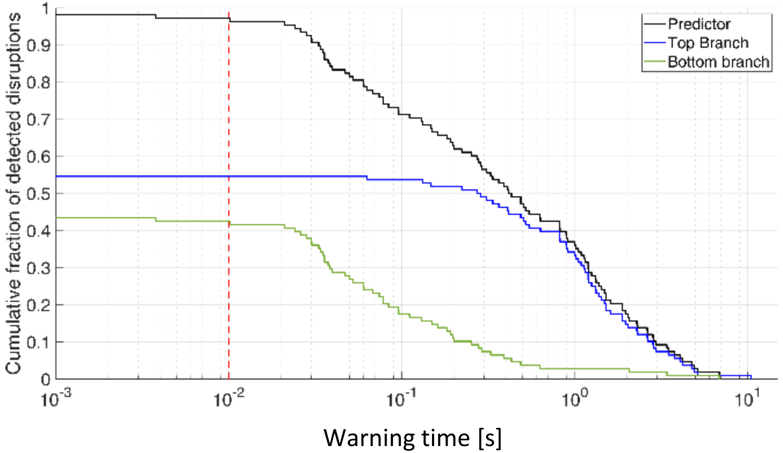




The separation of the two different mechanisms makes the predictor alarm more interpretable:

- top CNN branch provides larger warning times
- bottom CNN branch detects the mode-locking phase.

set	Campaigns	Disruptions	Regular
Training	C28-C30	85	70
Test	C36,C38	108	149



The predictor allows to greatly reduce the number of FAs

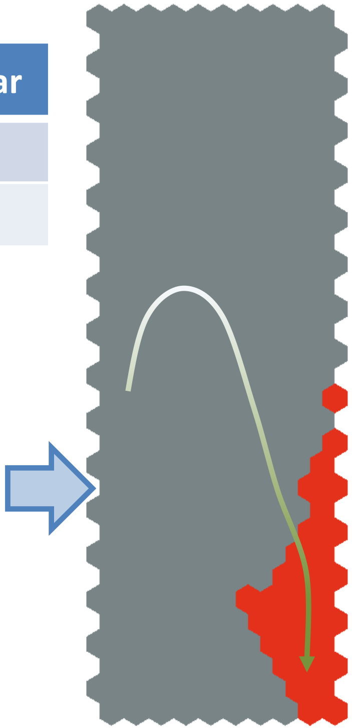
Performance index	CNN-UP1
MA-test [%]	1.87
FA-test [%]	0.67
Feature extraction	Automatic
Interpretability	Black box

SOM predictor: an unsupervised approach [6]

The SOM resulting from the unsupervised training is coloured providing it only with the information related to the discharge ending state: ***regular or disrupted***. No information about the precursors phase has been exploited.

set	Campaigns	Disruptions	Regular
Training	C28-C30	85	70
Test	C36,C38	108	149

Peaking factor of temperature	Te_{pf}
Peaking factor of electron density	ne_{pf}
Peaking factor_1 of the radiation(*)	$RAD_{pf\text{-}CVA}$
Peaking factor_2 of the radiation(**)	$RAD_{pf\text{-}XDIV}$
Internal Inductance	Li
Normalized locked mode	LM_{norm}



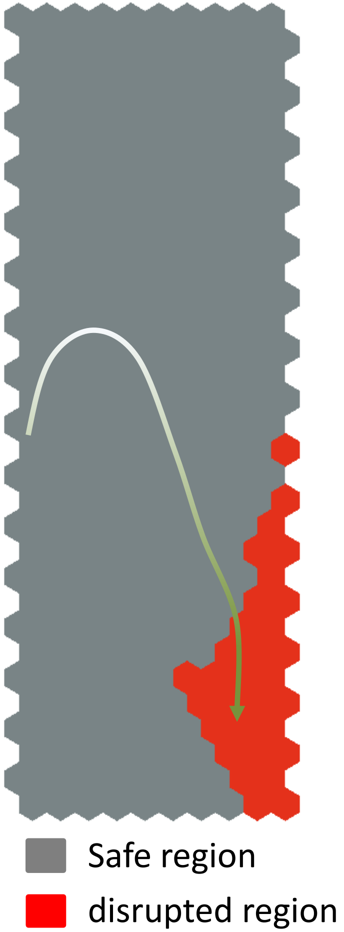
samples from disrupted and regular pulses

samples from disrupted pulses

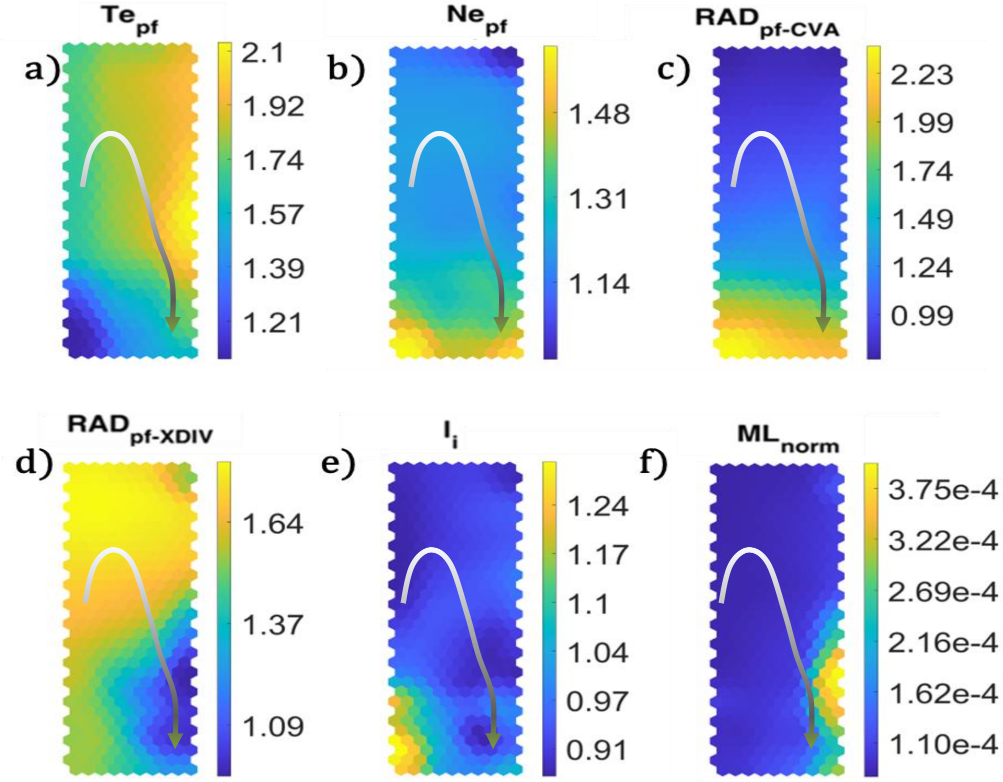
Performance index	SOM
MA-test [%]	4.63
FA-test [%]	2.01
Feature extraction	Manual
Interpretability	Yes

[6] E. Aymerich *et al.*, A self-organised partition of the high dimensional plasma parameter space for disruption prediction, **accepted for publication on Nucl. Fusion.**

SOM predictor: an unsupervised approach [6]



The evolution of the pulse can be tracked in real time while the monitoring the velues of the original variable on the SOM component plains



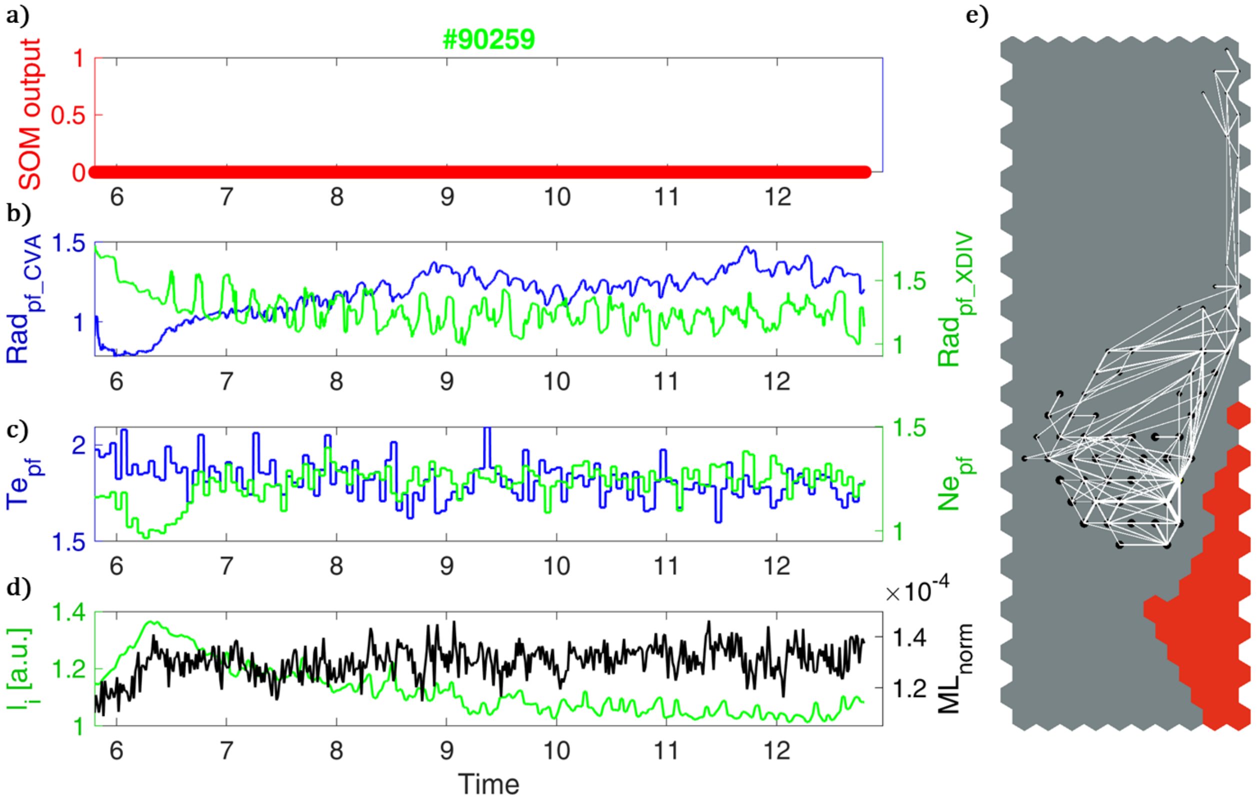
SOM predictor: an unsupervised approach [6]

The black dots track the position of the experiment on the map:

- beginning of the discharge flat-top
- ending of discharge flat-top

Regular pulse:

Flat and regular signal behaviors

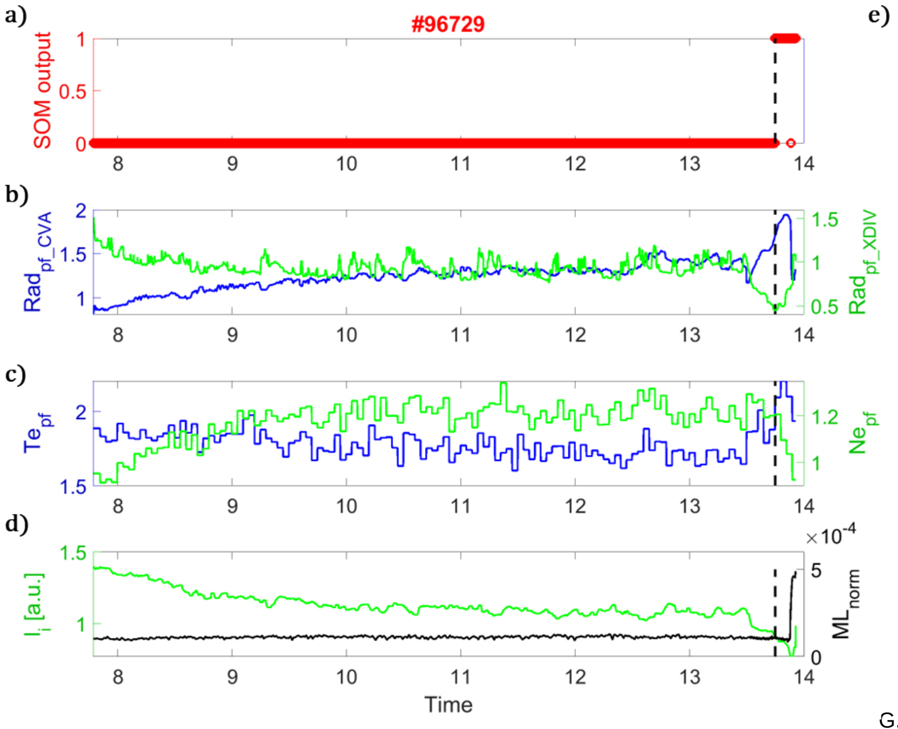


SOM predictor: an unsupervised approach [6]

The black dots track the position of the experiment on the map:

- beginning of the discharge flat-top
- ending of discharge flat-top

ending sample



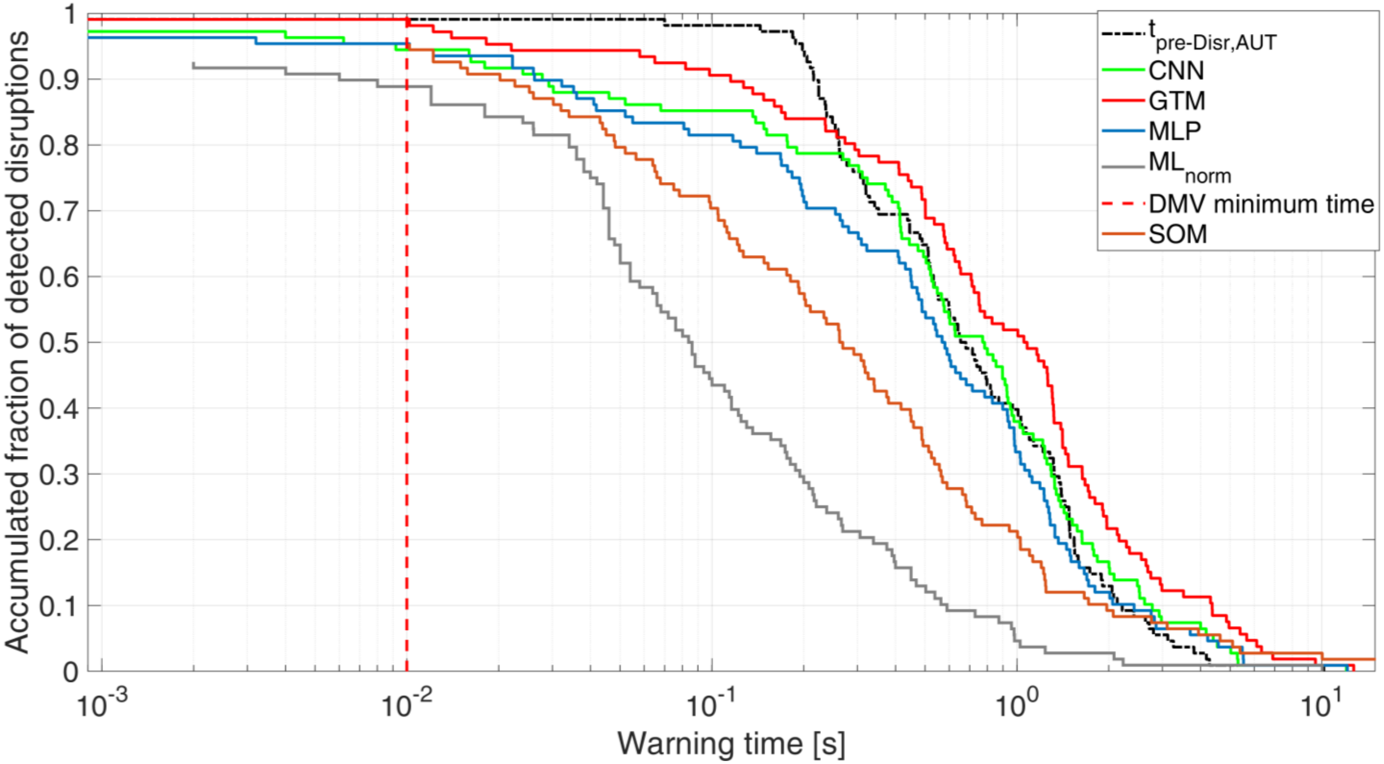
Transition from safe to the disruptive region:

- ✓ Increase of core radiation (increase of the $Rad_{pf\text{-}CVA}$ and decrease of the $Rad_{pf\text{-}XDIV}$, Figure 4b)
- ✓ Subsequent decrease of core temperature

Last disrupted phase

- ✓ Rise of the locked mode (ML_{norm})

Predictor performance comparison



Predictor	FA-test [%]
MLP	3.36
GTM	18.79
CNN	5.37
SOM	2.01

Conclusions and future works

- CNN predictor achieves good performance and doesn't need any preprocessing of plasma profiles (ne, Te and Prad)
  - ✓ Good a posteriori interpretability of the predictor answer for extrapolation of safe and disruptive path behaviors
  - ✓ Easy portability of the predictor to different machines after rescaling the plasma profile with respect the machine dimensions.

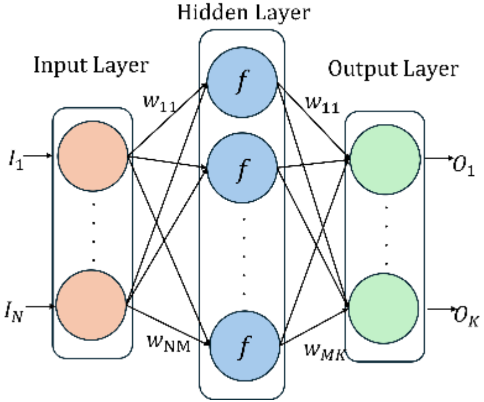
- SOM predictor achieved good performance with unsupervised training
  - ✓ No precursor phase is defined to interpret the predictor outcomes
  - ✓ Real-time tracking of the discharge on the map
  - ✓ Straight relation between the operative point evolution and features of the map regions for disruption monitoring.

- Ongoing work
  - ✓ Developing profile standardization algorithms for predictor portability
  - ✓ extracting rules from the SOM for a clear interpretation of the model's decisions during the discharge evolution.

Thank you

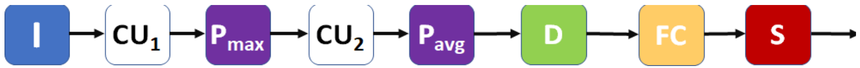
Neural networs

MLP models complex relationships between the input variable space $\bar{I}$ and the output variable space $\bar{O}$



$$
\begin{array}{l}
\text{Input Layer} \\
\text{Hidden Layer} \\
\text{Output Layer}
\end{array}
\begin{cases}
\bar{\bar{W}}_1 \cdot \bar{I} + \bar{b}_1 = \bar{g} \\
\bar{h} = f(\bar{g}) \\
\bar{\bar{W}}_2 \cdot \bar{h} + \bar{b}_2 = \bar{O}
\end{cases}
$$

CNN consists of a cascade of blocks which performs a filtering of an input image to extract significant features



convolutional unit CU_k $\boxed{CU_k} \equiv \boxed{C_k \mid N_k \mid A_k}$

- C_k convolutional layer
- N_k batch-normalization layer
- A_k nonlinear activation layer, with ReLU functions

P — P_{max} or P_{avg} are the max and average pooling layers

D — dropout layer

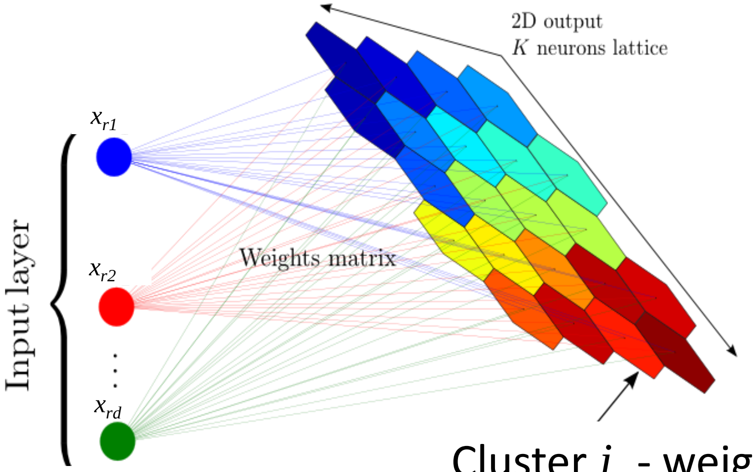
FC — Fully connected layer

S — SoftMax function

Self Organizing Map

A SOM projects a set of N d-dimensional input data $\mathbf{x}=[x_1, x_2, \ldots, x_d]$ into a $2D$ **discrete map topologically ordered**

Each input **x** is associated to a cluster of the map characterized by a weight vector **w** (barycenter of the inputs mapped in the node)



Cluster i - weight vector

$\mathbf{x_r}=[x_{r1}, x_{r2}, \ldots, x_{rd}]$

$r = 1, \ldots, N$ samples

$\mathbf{w_i}=[w_{i1}, w_{i2}, \ldots, w_{id}]$

How the SOM works

- ❑ **Competition**

find the winning neuron, i.e., the closest to each input vector

- ❑ **Cooperation**

find the winning neuron’s neighbors

- ❑ **Adaptation**

update the weights of winning neuron and its neighbors

$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + a h_{ij} [d(\mathbf{x}, \mathbf{w}_j(n))]$

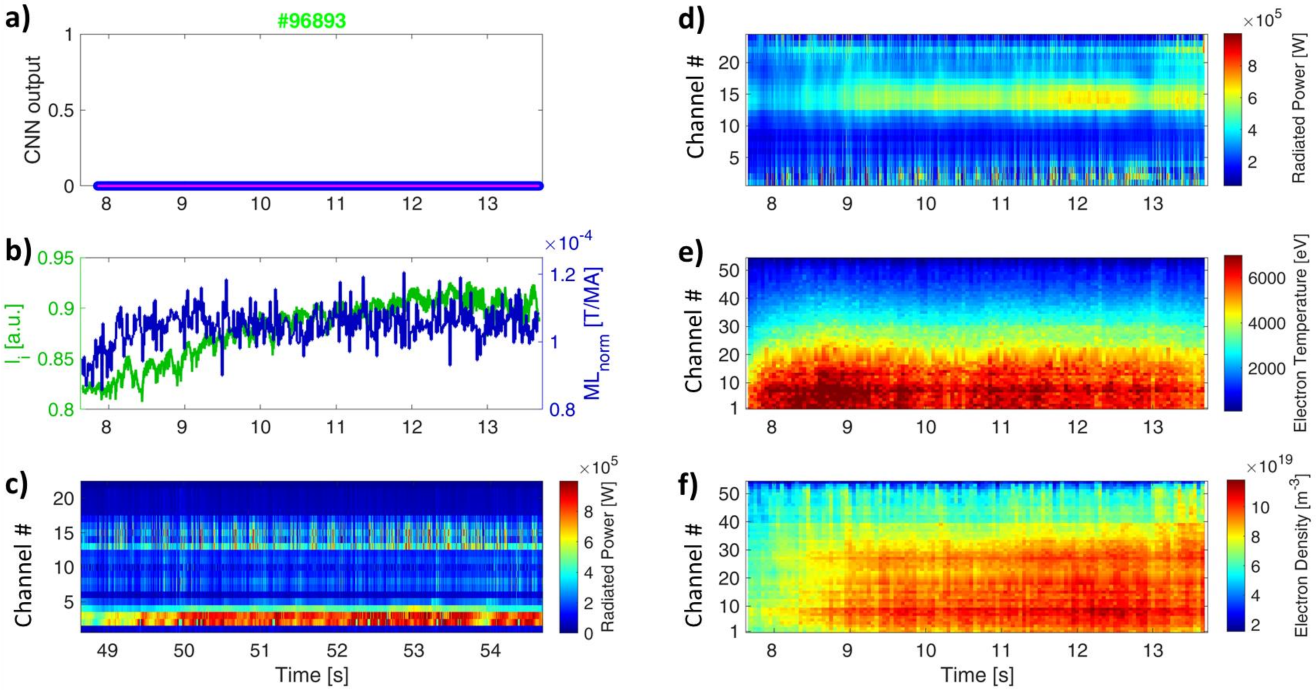
α learning rate

d distance function

h is the neighborhood function, it defines the winner neighborood

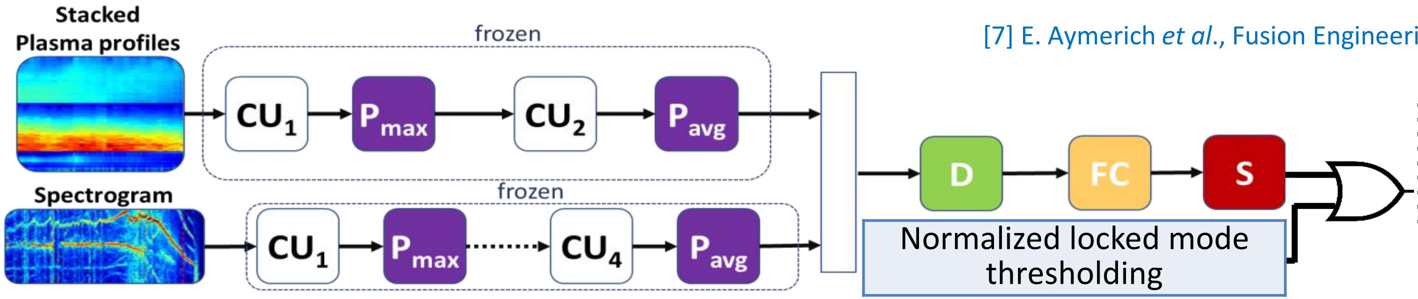
CNN predictor upgrade adding vertical bolometer data [5]

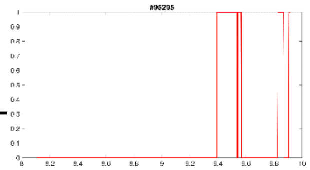
96893 is a regular pulse detected as disruptive by the CNN reference predictor [2], CNN-UP1 does not trigger an alarm, because the radiation pattern at chords #13-16 of BOL-H does not correspond to a radiation pattern of BOL-V.



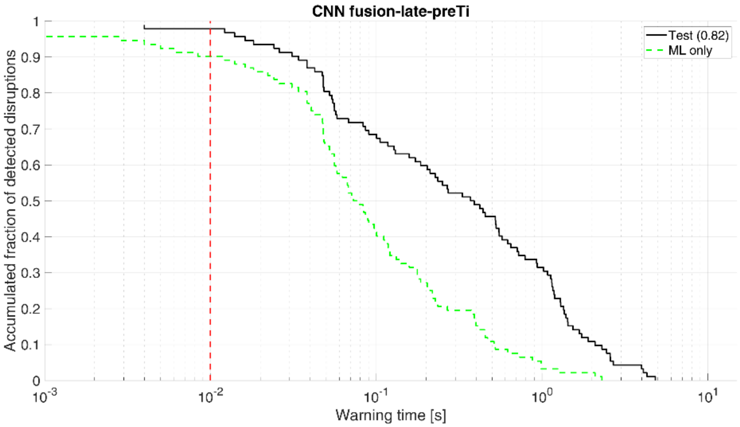
CNN predictor upgrade with MHD spectrogram [7]

[7] E. Aymerich *et al.*, Fusion Engineering and Design 204 (2024) 114472.





set	Campaigns	Disruptions	Regular
Training	C28-C30	75	65
Test	C36,C38	92	131



Remarkable reduction of FAs

Performancei Index	CNN-UP2
MA-test [%]	1.09
FA-test [%]	1.09
Feature extraction	Automatic
Interpretability	black box

CNN, responsible for processing the plasma profiles and Mirnov coils data, can yield longer warning times than the LM thresholding

CNN predictor upgrade with MHD spectrogram [7]

FA pattern in CNN reference predictor [2]:
- ✓ high radiation from central chords of BOL-H (figure 3d)
- ✓ decrease core electron temperature figure (figure 3e)
- ✓ peaking of the electron density at the core (figure 3f).

By adding the MHD spectrogram as input the CNN-UP2 output provides a limited rise of the disruptive likelihood both in time and value(figure 4a) with respect to the CNN reference predictor (figure 3a) .

CNN reference predictor [2]

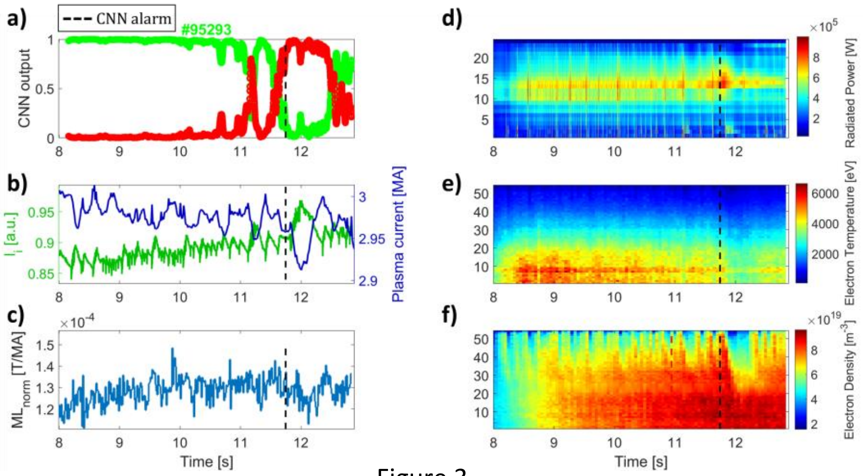


Figure 3

CNN-UP2 [6]

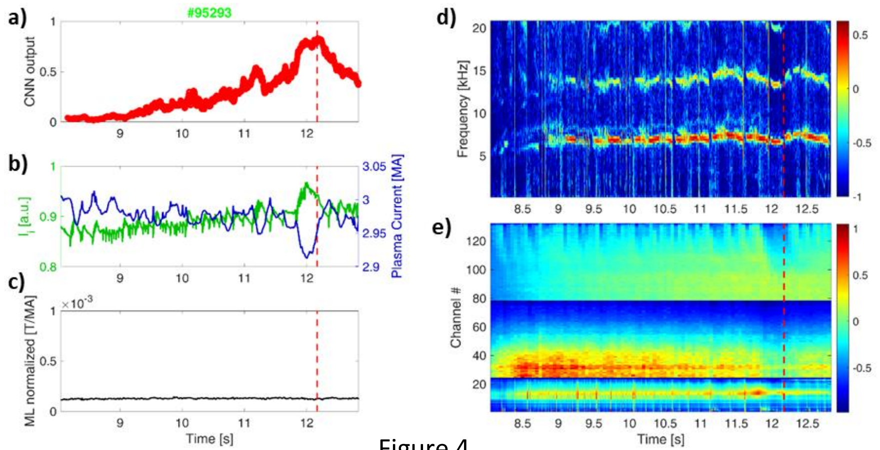
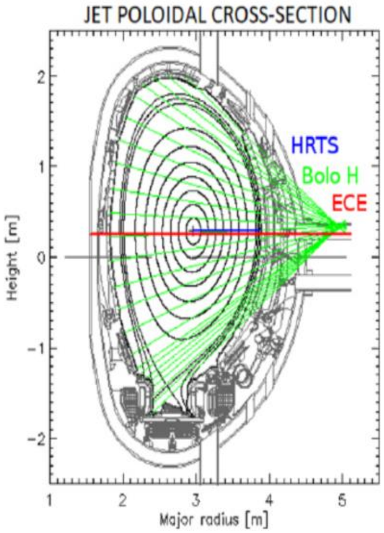


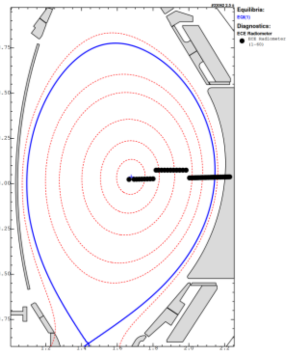
Figure 4

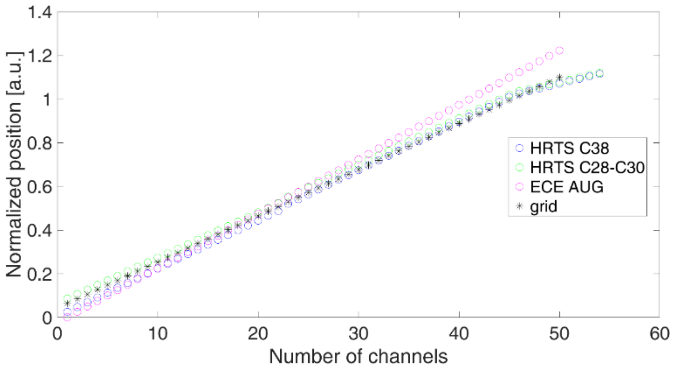
Profile standardization

Definition of ***resampling grids*** to standardize the profile images among JET and AUG machines



AUG - ECE

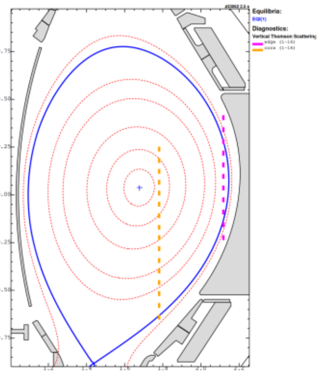


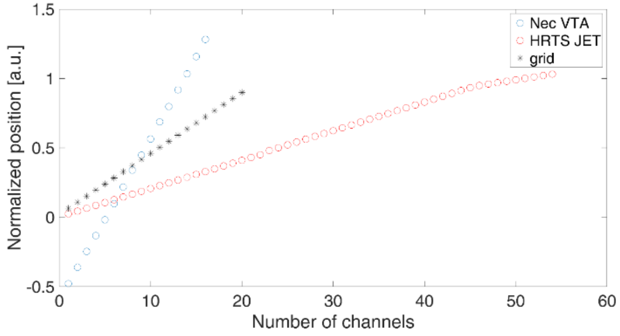


LoS normalized positions [a.u.]

$$r_i = \frac{X_i - R}{a}$$

AUG - VTA





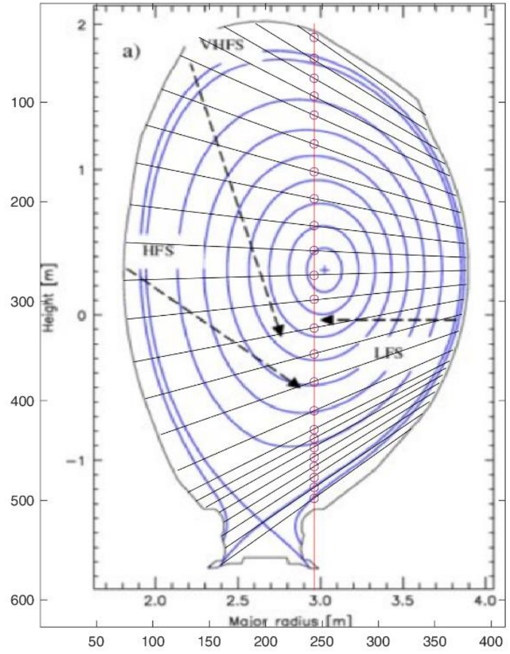
LoS normalized positions [a.u.]

$$r_{JET} = \frac{X - R}{a}$$

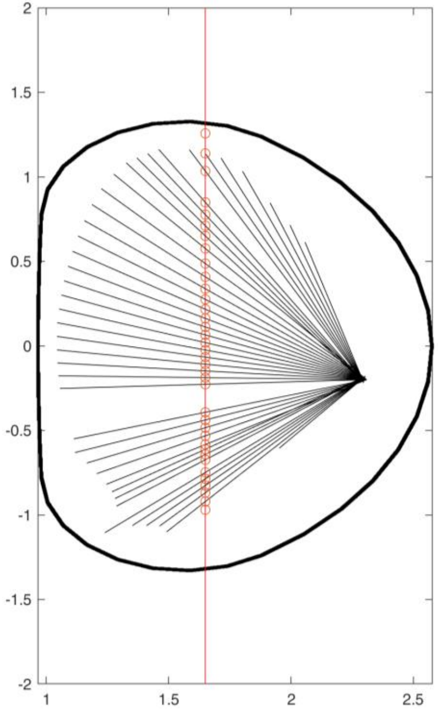
$$r_{AUG} = \frac{Z}{a}$$

Profile standardization

JET KB5-H geometry



AUG FHC geometry



LoS normalized positions [a.u.]

$$z_i = \frac{Z_i}{Z_{max} - Z_{\min}}$$