IAEA nuclear fusion

A self-organised partition of the high dimensional plasma parameter space for plasma disruption prediction

To cite this article: Enrico Aymerich *et al* 2024 *Nucl. Fusion* **64** 106063

OPEN ACCESS
IOP Publishing | International Atomic Energy Agency Nuclear Fusion
Nucl. Fusion 64 (2024) 106063 (15pp) https://doi.org/10.1088/1741-4326/ad7474

A self-organised partition of the high dimensional plasma parameter space for plasma disruption prediction

Enrico Aymerich* , Alessandra Fanni , Fabio Pisano ,
Giuliana Sias , Barbara Cannas , JET Contributors[a] and WPTE Team[b]

Department of Electrical and Electronic Engineering, University of Cagliari, Via Marengo, 3, Cagliari, CA 09123, Italy

E-mail: enrico.aymerich@unica.it

Received 14 May 2024, revised 7 August 2024
Accepted for publication 28 August 2024
Published 17 September 2024

Abstract

This paper introduces a disruption predictor constructed through a fully unsupervised two-dimensional mapping of the high-dimensional JET operational space. The primary strength of this disruption predictor lies in its inherent self-organization capability. Diverging from both supervised disruption predictors and earlier approaches suggested by the same authors, which were based on unsupervised models such as Self-Organizing or Generative Topographic Maps, this predictor eliminates the need for labeling data of disruption terminated pulses during training. In prior methods, labels were indeed required post-mapping to inform the model about the presence or absence of disruption precursors at each time instant during the disrupted discharges. In contrast, our approach in this study involves no labeling of data from disruption-terminated experiments. The Self-Organizing Map, operating without any a priori information, adeptly identifies the regions characterizing the pre-disruptive phase. Moreover, SOM discovers non-trivial relationships and captures the complicated interplay of device diagnostics on the internal plasma states from the experimental data. The provided model is highly interpretable; it allows the visualization of high-dimensional data and facilitates easy interrogation of the model to understand the reasons behind its correlations. Hence, utilizing SOMs across various devices can prove invaluable in extracting rules and identifying common patterns, thereby facilitating extrapolation to ITER of the knowledge acquired from existing tokamaks.

Keywords: self-organized map, disruption prediction and avoidance, interpretable machine learning, JET

(Some figures may appear in colour only in the online journal)

[a] See Maggi *et al* 2024 (https://doi.org/10.1088/1741-4326/ad3e16) for JET Contributors.
[b] See Joffrin *et al* 2024 (https://doi.org/10.1088/1741-4326/ad2be4) for the EUROfusion Tokamak Exploitation Team.
* Author to whom any correspondence should be addressed.

1. Introduction

Tokamaks, the most viable configuration for future fusion reactors, are prone to various types of instabilities. Disruptions are large scale plasma instabilities that cause a fast dissipation of the plasma's thermal and magnetic energy into the surrounding vessel and structure of the machine causing abrupt termination of discharges and material degradation. There is not a comprehensive theoretical model capable of reliably describing all types of disruptions. For this reason, data-based models utilizing machine learning (ML) are a common approach for classifying and predicting disruptions.

In general, the ML disruption predictors suffer from some drawbacks. Foremost, the models lack interpretability in terms of plasma dynamics. They require a large amount of data to be trained and they are typically limited to a specific tokamak. However, future reactors, with much higher stored energy, cannot provide enough unmitigated Disruption-Terminated Experiments (DTEs) at high performance to train the predictor without severely damaging the device. In the last decades, there have been few attempts to design disruption predictors, trained with experiments from a given device, that work also on different machines. This was done initially for JET and AUG [1, 2], while, more recently, some works focused on other cross-device predictors including DIII-D and EAST, especially using deep learning [3, 4].

Moreover, the low interpretability of neural network models is another drawback when considering their implementation on a real time control system for a critical application. For this reason Explainable AI (XAI) algorithms such as the ones in [5–7] had been applied to improve the interpretability of deep learning models. In [5], Class Activation Mapping has been adopted to understand the part of the image which was determining the deep learning algorithm decision, while in [6] other algorithms such as the occlusion and saliency maps are used for a similar purpose, and then an analysis to link the region of the image to the destabilizing mechanism (edge cooling or impurity accumulation) has been done. Instead the SHAP analysis adopted in [7] has a solid theoretical background in game theory, and it allows to estimate the feature contribution of the input features. The Shapley value is the contribution of a feature value to the difference between the actual prediction and the mean prediction when given the current set of feature values [7]. Instead, other approaches are with employ ML algorithms which are easily interpretable [8, 9]. In this case, the unsupervised SOM mapping maps N-D samples in a 2D map that can be visualized related to the plasma state values. Moreover, the map maintains the topological properties of the data space, so that close point samples in the N-D space will be close also in the latent 2D space. This means that, during the projection on the SOM, the cluster where a sample is projected gives us information on the data properties of the sample and of its neighborhood. Therefore, interpreting the SOM we can understand both the input influence on the model output and the properties of the data space itself.

Finally, there is an increasing integration of physics-based parameters together with physics laws in the predictors [7, 10–12], to increase the feature interpretability and guide the classification task towards the physics mechanism identification.

Most proposed ML models are supervised models, which require labeled training data. The manual labeling of the training data of DTEs, to identify the appearance of disruption precursors, is a heavy and challenging task. This problem has, in most cases, been addressed by assuming a temporal instant (called here as $t_{\text{pre-disr}}$) that identifies the pre-disruptive phase as equal for all disruptions [13, 14]. However, this is not consistent for all the experiments as different types of disruptions also have different precursor times. To overcome this limitation, in [15] a statistical algorithm capable of identifying a different pre-disruptive phase for each disrupted discharge has been proposed.

Few contributions are present where unsupervised methods are applied to the disruption prediction model. In this case the labeling is not necessary, and the model discovers by itself similarities and differences between the inputs. The works in [16, 17], presented disruption prediction methods based on anomaly detection. Training the anomaly detection models only required data from Regularly Terminated Experiments (RTEs) labeling them as 'normal'. When the model infers, data belonging to the pre-disruptive phase are classified as anomalous points, and a disruption alarm is issued.

This paper proposes an unsupervised disruption predictor for JET, based on Self-Organizing Maps (SOM). In previous predictors proposed by the same authors based on unsupervised models, such as SOM and Generative Topographic Mapping (GTM) ([9, 15, 18–21]), the maps were constructed using unlabeled data, even though the labeling of data from DTEs was still necessary for deploying the model as a disruption predictor. Conversely, in the present application, the SOM identifies the region where the pre-disruptive phase can be defined without assuming any *a priori* information. Moreover, the SOM allows one to visualize the high-dimensional plasma parameter space as a 2D projection. The obtained model is highly interpretable, and it can be easily interrogated to understand the reasoning behind the predictor answer with a closer connection with physics mechanisms. So, SOMs of different devices could be a valuable help for shared rule extraction and identification of common patterns towards a more confident extrapolation to ITER.

The paper is organized as follows: in section 2 the database used to train and test the model is described. The SOM and the selected performance indices are described in sections 3 and 4, respectively. Section 5 reports the explanation of the model's rationale. The results on disruption prediction are reported in section 6 and discussed in section 7 together with the alternative graphical representations of the SOM, such as the Component Planes and the Unified distance matrix. ln section 8 the conclusions are drawn.

Table 1. Database composition.

Dataset	Campaigns	Disruption Terminated Experiments (DTEs)	Regularly Terminated Experiments (RTEs)
1	2011–2013	127	115
2	2016	29	41
3	2019–2020	37	63

Table 2. Diagnostic signals, acronyms, and source diagnostics.

Plasma signal	Acronym	Diagnostics
Peaking factor of electron temperature	$\mathrm{Te_{pf}}$	HRTS
Peaking factor of electron density	$\mathrm{Ne_{pf}}$	HRTS
Peaking factor of the radiation (excluding the X-point/divertor region)	$\mathrm{Rad_{pf\text{-}CVA}}$	Bolometer horizontal camera
Peaking factor of the radiation (excluding the core region)	$\mathrm{Rad_{pf\text{-}XDIV}}$	Bolometer horizontal camera
Internal inductance	l_i	Magnetic equilibrium
Normalized locked mode amplitude	ML_{norm}	Saddle loops

2. Database

The data for this study comes from a database created and maintained by the University of Cagliari [15, 22], containing hundreds of DTEs and RTEs coming from several JET experimental campaigns, after the installation of the ITER-Like Wall (ILW), from 2011 to 2020. The considered database covers a wide set of experimental conditions, starting from the earlier campaigns with the ILW until the recent experiments where high power experiments were carried out. It has been grouped into three datasets, as detailed in table 1, following the different experimental campaigns.

In total, the database for this work contains a total of 193 DTEs and 219 RTEs having a flat-top plasma current higher than 1.5 MA, all diagnostic signals available and a flat-top length greater than 200 ms. Disruptions caused by Vertical Displacement Events have been excluded at all from the data set. Both flat-top and ramp down disruptions where the plasma current is over 1.5 MA are included in the dataset. These criteria are widely employed in disruption prediction and avoidance studies to select relevant experiments [4, 9]. The flat-top starting time has been assumed as the first time instant where the plasma is in X-point configuration. Both diagnostic and synthetic signals, derived from 1D plasma profiles, have been collected and they are listed in table 2.

The literature demonstrated the beneficial impact of the recent introduction of 1D plasma profiles [8–11, 15, 18, 22] as input to disruption predictors. In this paper, the temperature and density profiles come from the High-Resolution Thompson Scattering (HRTS), the Radiated Power profile comes from the horizontal lines of sight of the Bolometer, the internal inductance comes from the EFIT Magnetic equilibrium code, and the Locked Mode Amplitude comes from the Saddle Loops and is normalized by the plasma current.

The spatial information contained in the profile data has been synthesized by defining suitable 0D peaking factors as proposed in [9, 10].

The SOM has been trained and validated using a part of the dataset 1, by selecting the same 85 DTEs and 70 RTEs used in [8]. The remaining pulses of dataset 1 and all the pulses of datasets 2 and 3, resulting in 108 DTEs and 149 RTEs, have been used for testing the model performance and studying its behavior with unseen data also belonging to successive experimental campaigns.

Table 3. Composition of Training, Validation and Test sets.

Sets	DTEs	RTEs	JET campaigns
Training/Validation set	85	70	2011–2013
Test set	149	108	2011–2020

It is worth noting that dataset 3, which is related to experiments aiming to study the baseline scenario suitable for sustained high D–T fusion power, is characterized by higher currents, density, and input power, also exceeding the range of the other two datasets [5].

Table 3 reports the number of discharges and the originating campaigns, for the training/validation, and test sets.

The signals are available with a common time-base and sampled every 2 ms. Each sample consists in a 6 × 1 vector, where 6 is the number of signals in table 2. Training signals have been under-sampled (except for the final time interval of the DTEs) to limit map dimensions. Different values of the under-sampling and of DTE final time interval length have been tested to optimize the performance while maintaining a compact map representation. The validation of the SOM is carried out by feeding it with the same shots used for training with a sampling time of 2 ms. Subsequently, the performance of the SOM as disruption predictor is assessed using the independent test set outlined in table 3. This test set encompasses experiments from both the same campaigns as the training set and experiments from subsequent campaigns, as detailed in table 3.

3. Self-organizing maps

The SOM is a type of artificial neural network developed by Kohonen [23, 24]. It converts complex, nonlinear, statistical relationships between high-dimensional data items into

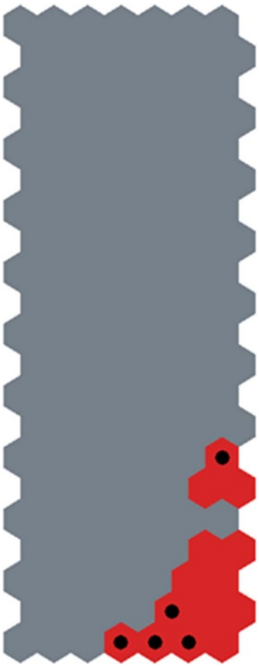

Figure 1. SOM-1. Red: clusters containing only samples from DTEs; grey: clusters containing samples from RTEs and DTEs. Black dots mark the disruptive clusters where the trajectories of the RTEs of the test set intersects them, triggering false alarms (FAs).

simple geometric relationships on a low-dimensional display. A SOM defines a mapping from the N-dimensional input space X ($N = 6$ in our case) onto a regular (usually two-dimensional) array of artificial neurons, preserving the topological properties of the input. This means that points close to each other in the input space are mapped on the same neuron or on neighboring neurons in the output space, i.e., the map is topologically ordered. Moreover, SOM also realizes the clustering of the input data because similar inputs will be mapped on the same neuron. The most common graphical representation of a SOM is the node map, which is a 2D grid or lattice, where each node represents a neuron formed by the SOM (see figure 1 as an example). The grid can be either rectangular or hexagonal, affecting how neurons interact with their neighbours. We used a hexagonal grid for this study. Each neuron corresponds to an N-dimensional weight vector, weight vector (called centroid) with the same dimensionality as the input data, and these weight vectors are adjusted during training to reflect the characteristics of the input data.

These centroid weights are initialized randomly. During the learning phase, for each sample $\mathbf{x}$, the goal is to determine the neuron, known as the Best Matching Unit (BMU), whose centroid is the closest in terms of Euclidean distance (or another measure of similarity). Subsequently, the BMU and the neurons in its neighborhood are updated, i.e. their weight vectors move toward the sample $\mathbf{x}$. This process is repeated over many iterations until a stopping condition is reached. When training is completed, the weight vectors associated with each neuron define the partitioning of the multidimensional data and the position of each cell on the grid reflects the similarity of its weight vector to those of neighboring neurons, with close cells representing similar neurons. Since we are interested in the grouping of similar points $\mathbf{x}$ together in the same neuron, in the following, we refer to the neurons as clusters.

The dimensions of the map can affect the quality of the mapping that is achieved by the SOM. Thus, several SOMs have been created varying the dimension of a rectangular grid topology. Moreover, the dimension suggested by the heuristic rule proposed by [25] for determining the size of the SOM grid has been tested, i.e. $k = \beta \cdot n^{0.54}$, where n is the number of examples (i.e. the number of N-dimensional time samples in the training set, in our problem). Values of 0.2, 1 and 5 are used for the constant β, which correspond to a small, normal and a large SOM, respectively [26].

In this paper, other possible graph visualizations of a SOM are used, including the Unified distance matrix (U-matrix), the Pie-Chart, and the Component Plane (CP) [27, 28].

The U-matrix depicts the distances between the BMUs of adjacent neurons in a grayscale image. Dark regions, with low values of distance, represent parts of the SOM where the adjacent neurons are close to each other. Lighter parts, with high values of distance, represent parts of the SOM where neighbour neurons are far away from each other. Consequently, light parts of the U-matrix indicate macro clusters boundaries, while dark parts reveal macro clusters themselves.

Pie Charts provide an overview of the distribution and proportion of different neurons. Each pie chart segment represents a neuron, with the size of each segment reflecting the proportion of input time samples in that neuron.

Component Planes are separate graphs depicting the values of each input variable across the SOM. Each component plane displays the distribution of a specific feature's values across the SOM grid, revealing patterns and variations in the data.

The CPs provide a visual insight into how different features contribute to the organization of the input space within the map. They are displayed in various shades of colors or grey scale colors on the maps. From the observation of CPs, possible correlations between the input variables can be identified.

After training, each N-dimensional input time sample from any training discharges is assigned to the nearest neuron in the map and a range can be determined for each cluster in the map by the minimum and maximum signal values of the samples associated with that neuron. The SOM serves as a low-dimensional representation of the input parameter space, allowing new discharges to be projected onto it. Each original N-dimensional time sample from a discharge is mapped to the nearest cluster on the SOM, forming a trajectory.

4. Prediction performance indexes

Depending on the alarm time, the resulting warning time, which is the time interval between the alarm time (t_{alarm}) and the disruption time t_{D}, a different intervention can be put in place, such as active control, avoidance, or mitigation. If in any case the alarm is activated by the predictor with a warning time equal to or greater than the minimum activation time for the mitigation system, it is assumed as a successful prediction. If the warning time is not even sufficient for mitigation, the prediction is classified as Tardy Detection (TD). At JET, the minimum warning time is 10 ms, which is the time required to the Massive Gas Injection system (MGI) to mitigate the discharge. A Missed Alarm (MA) occurs if the disruption prediction system does not trigger any alarm. At JET, a conclusive definition of premature alarms has not yet been established, so in the following, premature detections will not be counted. Furthermore, false alarms (FA) must be considered when evaluating the performance of the disruption predictor. The FAs are alarms triggered by the predictor in response to a regularly terminated discharge.

In the disruption prediction literature, a most informative figure of merit is defined by the accumulated fraction of detected disruptions as a function of the warning time. It allows to read, in a unique graph, besides the successful prediction and the tardy detections, also a general overview of the premature detections and the alarm anticipation times.

5. Self-organized labeling of the plasma operative space through SOM

The operational space of the plasma described by the six plasma parameters in table 2, available for the 85 DTEs and 70 RTEs in table 3, was projected onto a 2D SOM. Note that, the SOM is inherently an unsupervised algorithm, meaning no information about the labeling of training samples is provided during training. The SOM obtained after this unsupervised training phase effectively captures the spatial organization of the data.

Unsupervised approaches to disruption prediction are also found in [8–10, 18–20]. In [18–20], some kind of knowledge was added to the SOM by means of a subsequent supervised phase consisting in coloring the clusters of the SOM using information on the length of pre-disruptive phase for disruption terminated experiments (DTEs). In [19, 20] a label (and consequently a color) is associated to each sample in the training set: a safe label was associated to a sample of an RTE or to a sample of a DTE not belonging to pre-disruptive phase; a disruptive label was associated to a sample of a DTE belonging to the previously defined pre-disruptive phase. To this end, a time instant $\mathrm{t}_{\mathrm{pre\text{-}disr}}$ was defined for the DTEs, which discriminates between the non-disruptive and the pre-disruptive phases. The identification of the pre-disruptive phase is anything but simple, and only an estimation can be made, introducing uncertainties, nonetheless. In [7], a statistical procedure was developed that achieved results consistent with a manual evaluation based on physics. In [9, 10] a GTM map of the JET operational space was obtained by using the algorithm in [7] to identify the time instant $t_{\mathrm{pre\text{-}disr}}$.

In our paper, the SOM resulting from the unsupervised training, was then colored providing it only with the information related to the discharge ending state: regular or disrupted. Note that, the distinction among samples belonging to the stable phase of DTEs and samples with a high risk of disruption (belonging to the pre-disruptive phase) is carried out by the SOM itself without giving it any further information. To proceed to the SOM coloring, let us firstly define the label for every cluster. A cluster is labeled as 'safe' or 'disruptive' depending on its composition:

- Safe clusters are all those clusters containing at least one sample from RTEs. In some of these clusters there are only samples from RTEs, in others there are both samples from RTEs and DTEs. Since these clusters contain RTEs samples, the associated disruption risk is considered to be low. So, the cluster is labeled as safe. In the node map representation, it is colored in grey. After coloring phase, the grey cluster could form macro-clusters identifying 'safe' regions (see figure 1, for example).
- Disruptive clusters consist solely of samples from DTEs. As these clusters lack samples from RTEs, they are regarded at high disruption risk, and in the node map representation, it is colored in red (see figure 1). During the evolution of a discharge, samples from DTEs may populate clusters containing samples from RTEs. However, as the disruption approaches, disruption precursors could start emerging and the high-dimensional operational space, as described by the DTEs samples, deviates from the space outlined by the RTEs samples. Operatively, owing to the self-organized nature of the method, as disruption approaches, the DTEs samples will populate disruptive clusters simply because they differ from samples of RTEs.

In the proposed approach, the sample label depends on the position of the sample in the 2D map, i.e. by the cluster associated to the sample. Indeed, the sample label is the label of its cluster. Note that, due to the Self-Organized nature of the approach, in principle, samples occurring before the appearance of disruption precursors, could be associated to disruptive clusters because they are simply different from samples of RTEs.

Table 4. Performance of SOM-1: (*a*) without novelty detection and (*b*) with novelty detection.

	(a): SOM-1 (without novelty)			(b): SOM-1 (with novelty)		
	FA	MA	TD	FA	MA	TD
Training	0%	1.18%	1.18%	0%	1.18%	1.18%
	(0/70)	1/85	1/85	(0/70)	1/85	1/85
Validation	0%	1.18%	1.18%	0%	1.18%	1.18%
	(0/70)	1/85	1/85	(0/70)	1/85	1/85
Test	4.70%	3.70%	4.63%	4.35%	4.63%	3.70%
	(7/149)	4/108	5/108	(6/149)	5/108	4/108

The different graphical representation of the produced SOM can be used both to visualize and analyze the operational space and to monitor the plasma state during a discharge simulating the on-line operation. In fact, as previously cited, the temporal sequence of the samples in a discharge can be projected on the SOM, obtaining a trajectory that describes the discharge evolution. When a sample leaves the safe region and reaches a disruptive cluster, the alarm should be triggered.

6. Results

The optimal model was selected by optimizing its performance on the same set of experiments used to train the map, but without under-sampling (validation set). Each discharge is then projected onto the map, and an alarm is activated if a sample is projected into a disruptive cluster. During this validation phase, maps with different number of clusters and neighborhood shape have been trained on training sets sampled with different sampling times $\Delta\tau$ and compared in terms of performance on the validation set (where, instead, the sampling time is always equal to 2 ms). The selected map has 21 × 7 hexagonal clusters, $\Delta\tau = 10$ ms, and the last second of DTEs sampled with $\Delta\tau = 2$ ms. The SOM shows only two errors in the validation set, while keeping a reduced number of clusters (147). The map size has been optimized by maximizing the performance on the validation set while minimizing the number k of SOM clusters, to reduce the risk of incurring in overfitting.

Figure 1 shows the node map representation of the obtained SOM (SOM-1); there are no clusters containing samples exclusively from RTEs, and most clusters contain samples from both RTEs and DTEs and are shaded in grey. This means that DTEs and RTEs generally start in the grey region of the plasma parameter space and differentiate their trajectories after a while. The red clusters in figure 1 consist exclusively of samples from DTEs.

The SOM can be used as disruption predictor by projecting a discharge onto it. During the experiment evolution, when discharge projection falls into a disruptive cluster, the alarm is triggered. Then, depending on the knowledge of the experiment outcome (DTE, RTE), the prediction is considered successful, or a TD, or a FA. Table 4(*a*) documents the quite good disruption prediction performance across the training, validation, and test sets. The prediction success rate stands at approximately 93% with a false alarm rate of less than 5%, evaluated on the test discharges that were never presented to the model during its training.

Figure 2 reports the cumulative fraction of disruptions detected by the predictor as a function of the warning time. The warning time represents the time interval between the predictor alarm time and the disruption time. An appropriately timed warning provides the control system with the opportunity to respond to the onset of instabilities. Conversely, in cases of a brief warning time, the disruption is typically managed at JET through the activation of the MGI. In figure 2, the green line depicts the SOM-1 warning time, assessed by considering all the DTEs in the test set. This singular graph offers a comprehensive overview, providing a holistic understanding of both premature detections and alarm anticipation times. Additionally, it aids in interpreting the successful prediction fraction (SP), denoting the intersection between the cumulative curve and the minimum anticipation time (10 ms required at JET to initiate mitigation actions, indicated by the red dashed vertical line). Detections occurring beyond this line are classified as delayed alarms. Consequently, the TD fraction is computed as 1-SP.

To interpret the SOM behavior and evaluate the reliability of its answers, the range of each cluster may be considered, i.e. the minimum and maximum values of the six plasma parameters for the training samples mapped inside the cluster. During the projection of a discharge, the generic test sample is mapped into a cluster based on proximity, but some of its signal values might fall outside the range of that cluster. Then, if at least one signal value is outside this range, with a tolerance of 10%, the sample is considered as novel. Table 4(*b*) reports the performance on training, validation and test sets evaluated when the disruption alarm is triggered only for not novel samples. In other words, it reports results after excluding all alarms triggered in correspondence to novel samples. In comparison to table 4(*a*), the introduction of novelty detection does not affect performance in the training and validation sets. However, in the test set, a tardy detection turns into a missed alarm, while keeping the correct predictions unchanged. Additionally, the occurrence of a false alarm is avoided. The blue curve in figure 2 shows the cumulative fraction of detected disruptions for the test set when novelty detection is considered. The green and blue curves in figure 2 do not deviate significantly, just as the results presented in tables 4(*a*) and (b), showcasing the robustness of the proposed model.

The frontier between safe and disrupted clusters is critical for the performance of the model. In figure 1, the black dots

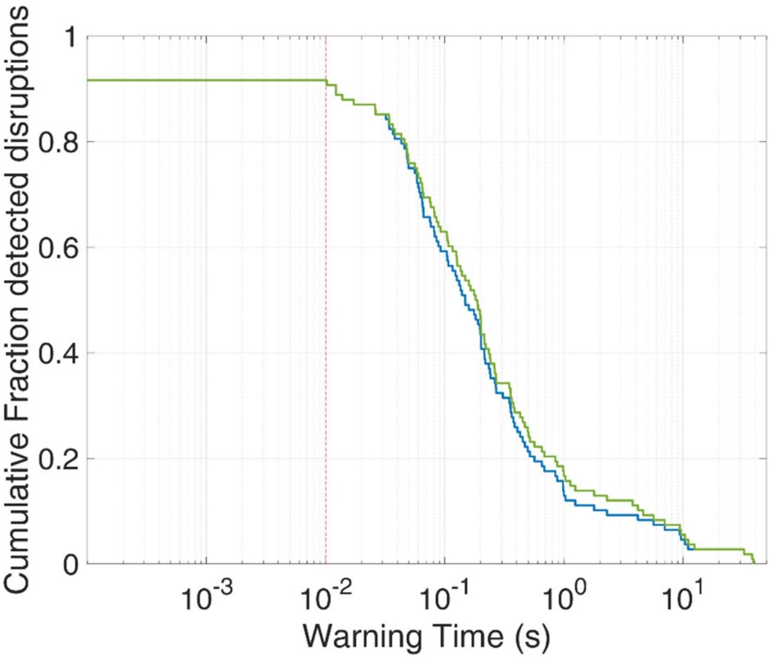

Figure 2. Cumulative fraction of disruptions detected by SOM-1 as a function of the warning time. Green curve refers to alarms triggered for all the DTEs in the test set without considering alarms activated by samples recognized as novel. The blue curve refers to alarms triggered for all DTEs in the test set, excluding those related to samples recognized as novel. The vertical red line indicates the minimum time necessary for mitigation actions at JET.

pinpoint the clusters in which test samples from RTEs, leading to false alarms, are projected. Looking at figure 1, most of the false alarms are triggered in disrupted clusters adjacent to the non-disruptive region. This is not surprising because, in the boundary region between safe and disruptive regions, samples of RTEs and DTEs are most similar; thus, the distinction between them is more intricated.

For this reason, another model with a larger map, 32 × 10, which kept a similar aspect ratio of SOM-1 was trained. This should allow us to stretch the boundary and better separate RTE and DTE plasma states. The obtained node map representation of the SOM (SOM-2_a, shown in figure 3, has the performance reported in table 5(*a*).

The disruption predictor has no missed or tardy alarms, but a large number of false alarms. The black dots pinpoint the clusters in which test samples from RTEs, leading to false alarms, are projected. As it can be noticed, FAs are mostly triggered on the frontier, similarly to SOM-1. To reduce false alarms, disruptive clusters adjacent to a non-disruptive cluster of SOM-2_a have been recolored in grey to provide a 'safety margin' before triggering an alarm. The node map representation of the resulting map, SOM-2_b, is shown in figure 4. The performance of SOM-2_b is shown in table 4(*b*). In this case, both performance is computed by deactivating the SOM alarm when the sample is out of range of the cluster where it is projected, i.e., if it is recognized as novel. By comparing the performance of SOM-2_b with that of SOM-1 in table 3, it is possible to see that SOM-2_b provides a lower number of overall errors, with a successful prediction rate greater than 95%, while simultaneously reducing the false and tardy alarms.

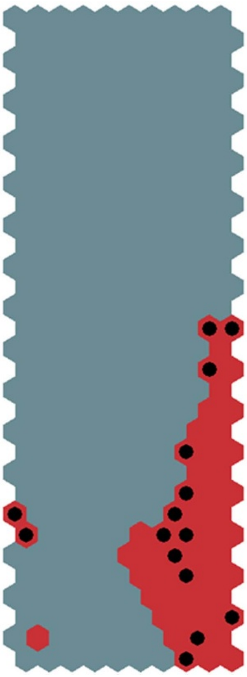

Figure 3. SOM-2_a. Red: clusters containing only samples from DTEs; grey: clusters containing samples from RTEs and DTEs. Black dots mark the disruptive clusters where false alarms (FAs) are triggered.

For the sake of comparison, table 6 reports the performance obtained by the same authors in [22] with a disruption predictor consisting of a Convolutional Neural Network (CNN) trained and tested on the same experiments. Whereas the SOM

Table 5. Performance of (*a*) SOM-2_a and *b*) SOM-2_b obtained coloring in grey the disruptive (red) clusters adjacent to non-disrupted grey clusters. The performance is obtained deactivating the model when the projection is out of cluster range (novelty criterion).

	(a) SOM-2_a			(b) SOM-2_b		
	FA	MA	TD	FA	MA	TD
Training	1.43% (1/70)	0% 0/85	1.18% 1/85	0% (0/70)	0% 0/85	1.18% 1/85
Validation	1.43% (1/70)	0% 0/85	1.18% 1/85	0% (0/70)	0% 0/85	1.18% 1/85
Test	51.01% (76/149)	0% (0/108)	0% 0/108	2.01% (3/149)	4.63% 5/108	0% 0/108

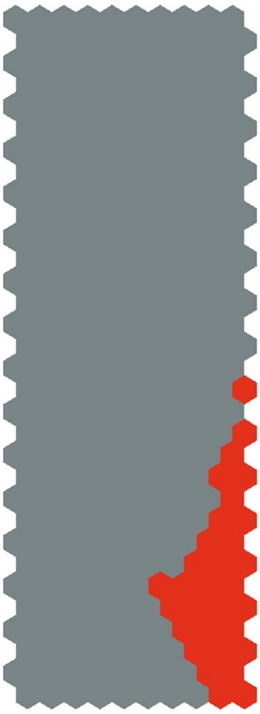

Figure 4. SOM-2_b obtained coloring in grey the disruptive (red) clusters adjacent to non-disruptive grey clusters in SOM-2_a.

has as input the temperature, density and plasma radiation profile peaking factors, the CNN directly processes the raw plasma profiles data that are converted into images and vertically stacked. Internal inductance and Locked Mode signals are supplied in input to both SOM and CNN predictors. As CNN is a supervised algorithm, during the training, a label was assigned to the time samples of the input plasma parameters. For each DTE, the labeling was carried out by automatically identifying the pre-disruptive phase by means of the algorithm proposed in [15].

As it can be noticed, despite a comparable number of MAs, SOM-2_b exhibits much better performance in terms of FAs. This is evidently owed to the conservative choice made for the SOM clusters labeling, i.e. considering as non-disruptive all the clusters where at least one sample of an RTE is present. This choice did not compromise performance in terms of disruptions detected in time for disruption mitigation. However, as shown by the cumulative warning time distribution reported in figure 5, for both SOM and CNN, the two curves are nearly superimposed until approximately 20 ms before the disruption time whereas, for warning times greater than 10 ms, CNN is able to trigger the alarms well in advance. Therefore, the CNN can provide alarms with a larger warning time than the SOM, which on the other hand has a much lower number of FAs.

7. SOM analysis

The SOM could help extract physics knowledge from plasma experimental data and bridge the gap between theoretical models and practical implementations. First of all, during the experiment, the current process state and its history in time could be visualized as a trajectory on the map, in order to monitor the plasma position and its closeness to the frontier between safe and disruptive region. At the same time, the SOM can supply the numerical information about the values

Table 6. CNN performance [22].

	CNN		
	FA	MA	TD
Training + Validation	4.3% (3/70)	0% 0/85	0% 0/85
Test	9.4% (14/149)	3.7% (4/108)	1.85% (2/108)

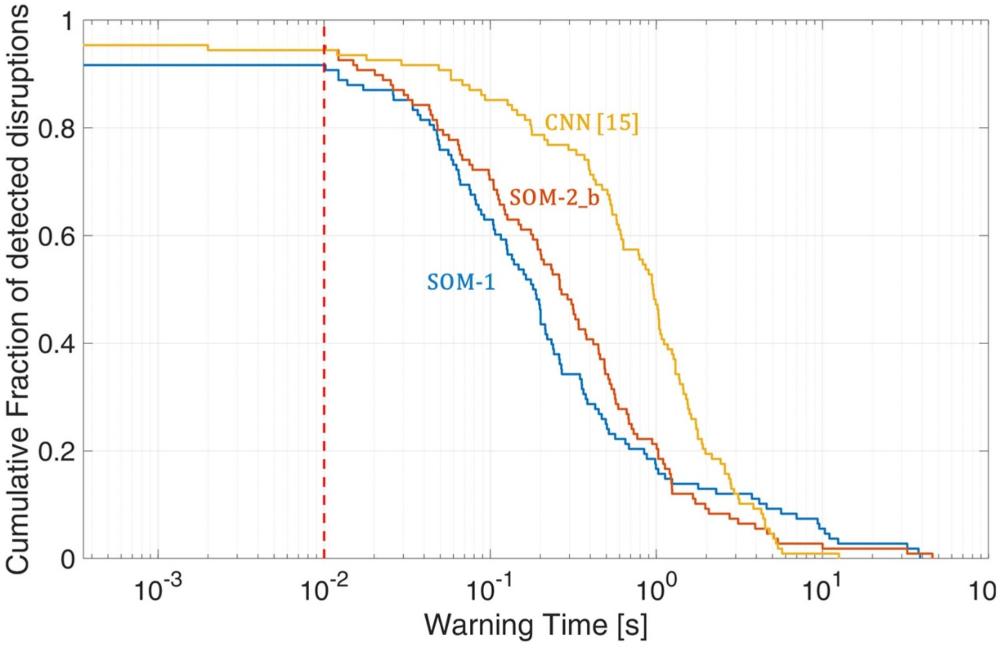

Figure 5. Cumulative fraction of disruptions detected by SOM-1 (blue), SOM-2_b (orange) and CNN [22] (yellow) predictors as a function of the warning time.

of the plasma parameters in the different regions of the operative space described during RTEs and DTEs. Figures 6 and 7 show the trajectories of the plasma operative point for the RTE 90 259 and the DTE 96 729 respectively. In the figure 6(*e*), the black dots track the position of the experiment on the map; the smaller dots represent the beginning of the discharge flat-top and the larger ones the end of the flat-top. It is possible to notice how the discharge starts in the top-right corner of the map and later gets to the middle of the map, without entering the disruptive (red) region. On the other hand, in figure 7(*e*), the DTE also starts in the grey area of the map, then moves closer to the frontier of the map and enters the disruptive area a first time. The frontier crossing is bordered with two red dots, and it is highlighted by a black dashed line in figures 7(*a*)–(d) (at 13.75 s). Then, the pulse goes back in a grey cluster to return in the disruptive region a second time just for one sample. From figure 8(*a*)) it can be noted that the grey cluster is populated mainly by disruptive samples. The ending sample of the experiment projection is marked by a yellow dot. It is possible to notice that the first transition in the disruptive region is characterized by an increase of the $\mathrm{RAD}_{\mathrm{PF\text{-}CVA}}$ values and a decrease of the $\mathrm{Rad}_{\mathrm{pf\text{-}XDIV}}$ ones which are shown in figure 7(*b*). Moreover, this increase of core radiation is followed by a later decrease of the $\mathrm{Te}_{\mathrm{pf}}$ signal, similarly to what observed in previous works with temperature hollowing [6, 9, 10, 29]. The last part of the evolution in the red area corresponds instead to the rise of the $\mathrm{ML}_{\mathrm{norm}}$ (after 13.88 s) which finalize the destabilization of the discharge.

Figure 8 shows the cluster composition of SOM-2_b in terms of training (figure 8(*a*)) and test data (figure 8(*b*)). In the pie charts superimposed to the clusters, green slices correspond to RTE samples and magenta slices to DTE samples. It can be noted that the RTEs time samples of the test set occupy the plasma parameter space differently, shifting closer to the boundary between non-disruptive and disruptive regions.

Figure 9 reports the SOM-2_b with the pie charts for RTEs (figure 9(*a*)) and DTEs (figure 9(*b*)) representing the data distribution during the different experimental campaigns in the test dataset (Dataset 1 in blue, Dataset 2 in green, Dataset 3 in yellow). It can be observed in figure 9(*a*) that, while DTEs fill the entire map during the first campaigns (2011–2013), the RTE ones are mainly projected onto the left, top-left and top of the map (figure 9(*b*)). In the latest campaigns (2016 and 2019–2020), the RTEs progressively moved towards the right and the

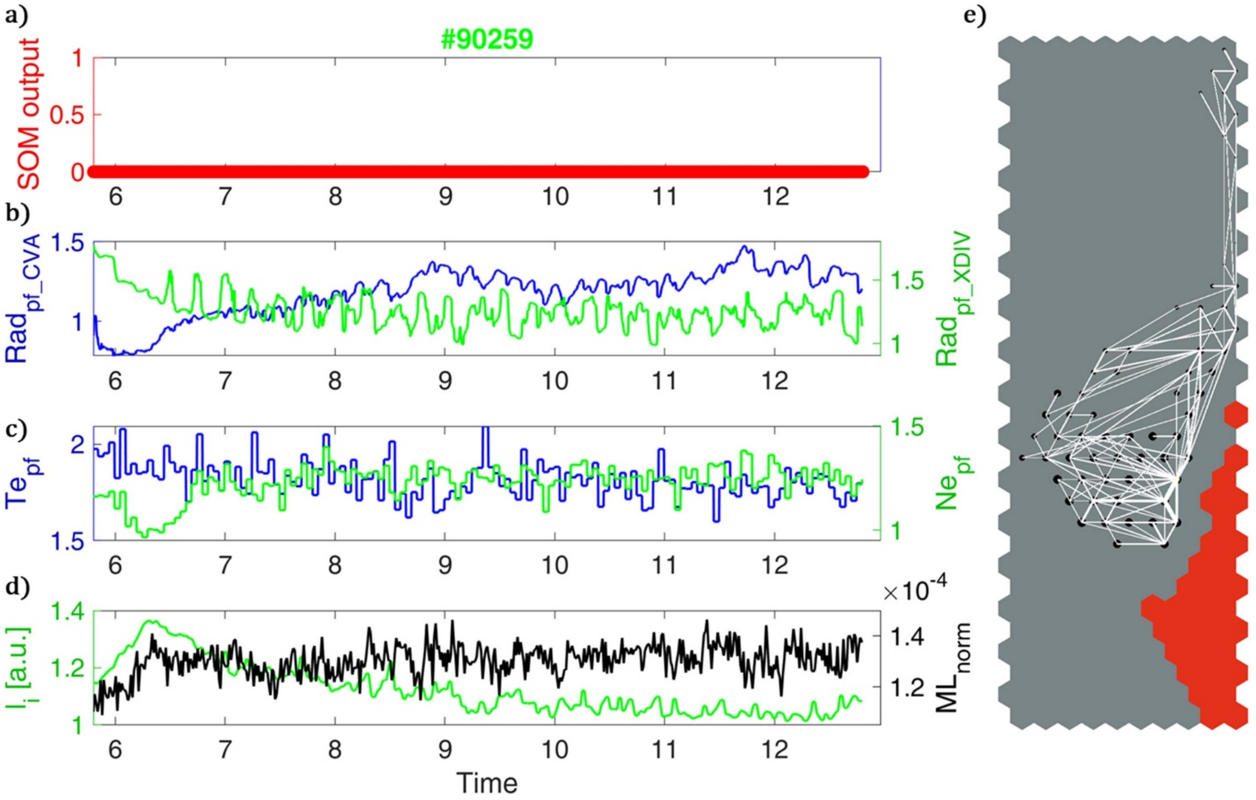

Figure 6. RTE Example 90 259 *(a)* SOM output where the 0 values represent no disruption alarm and 1 am alarm. *(b)* Rad$_{\text{pf-CVA}}$ (in blue) and Rad$_{\text{pf-XDIV}}$(in green) signals *(c)* Te$_{\text{pf}}$(in blue) and Ne$_{\text{pf}}$ (in green) signals *(d)* li (in blue) and ML$_{\text{norm}}$ (in black) signals *(e)* Trajectory of the experiment on the SOM-2_b, where the black dots mark the projection of the discharge. Smaller dots indicate the beginning of the flat-top, and larger dots the end of the flat-top.

lower part of the map, getting closer to the disruptive region. The experiments of the campaign in Dataset 3 do not occupy the top-left area anymore. Summarizing, figure 9 shows how the experimental conditions in the later campaigns push the discharges closer to the transition region. The increased risk of disruption is corroborated by an intensified intervention of soft stop systems during these campaigns [12, 30–32].

Figure 10 illustrates the U-Matrix representation of SOM-2_b, where each cell in the matrix represents the distance between neighboring neurons' weight vectors. The U-Matrix is grey-scale-coded to indicate these distances, with lighter colors signifying larger distances and darker colors indicating smaller distances. As can be noted, it exhibits two areas with a higher inter-cluster distance, one on the left side, in the grey area, and the other on the right side of the map. These regions are both associated with a high disruption risk as shown in figure 8(*a*). Thus, the transition from the safe to the disruptive region appears characterized by a consistent variation of the plasma parameters.

In the bottom left side of SOM-2_b there is a large grey region. As shown in figure 6, this area, despite being grey, is populated by a large number of samples coming from DTEs.

As previously cited, with the so-called Component planes, shown in figure 11, we can visualize the weights of the individual plasma parameters as 2D plots obtaining qualitative information about how the training input variables are related to each other. For example, we can suppose that the grey region on the lower part of the map is associated with impurity accumulation. In fact, the region is characterized by low values of the temperature peaking factor Te$_{\text{pf}}$, and high values of the density and radiation peaking factors Ne$_{\text{pf}}$ and RAD$_{\text{pf - CVA}}$.

On the other hand, on the right side of the SOM-2_b there is a compact red area. This area is characterized, as visible from the Component planes in figure 11, by low values of the RAD$_{\text{pf}-\text{XDIV}}$ peaking factor (11(d)) and high values of the ML$_{\text{norm}}$ (11(*f*)). The other parameters vary over this region, which do not represent a specific disruption class. For instance, in the bottom right part of the red area the pattern is similar to the bottom left one, with low values of the temperature peaking factor Te$_{\text{pf}}$, and high values of the density and radiation peaking factors Ne$_{\text{pf}}$ and RAD$_{\text{pf}-\text{CVA}}$. On the other hand, in the top of the red area there are high Te$_{\text{pf}}$ and Ne$_{\text{pf}}$ values.

Moreover, figure 12 reports the histogram of the ML$_{\text{norm}}$ values for the time samples of the test pulses which fall into the disruptive clusters (in blue) together with the ML_{norm} histogram of the BMUs of the disruptive clusters (in red). It is possible to see that both high and low values of the ML_{norm} are present in the disruptive clusters. Since low values of ML_{norm} characterize the centroids of the disruptive clusters, this means that these clusters present off-normal patterns in

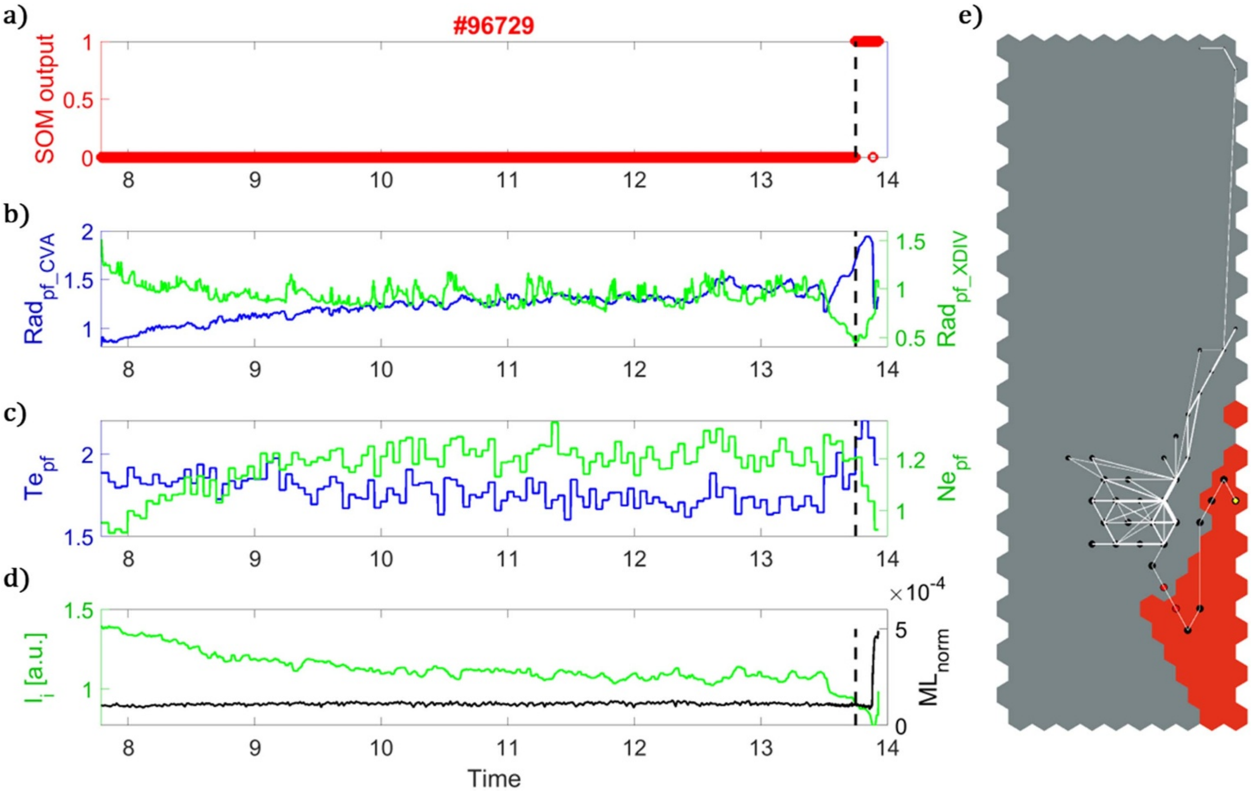

Figure 7. DTE Example 96 729 *(a)* SOM output where the 0 values represent no disruption alarm and 1 am alarm. *(b)* $\mathrm{Rad}_{\mathrm{pf\text{-}CVA}}$ (in blue) and $\mathrm{Rad}_{\mathrm{pf\text{-}XDIV}}$ (in green) signals *(c)* $\mathrm{Te}_{\mathrm{pf}}$ (in blue) and $\mathrm{Ne}_{\mathrm{pf}}$ (in green) signals *(d)* li (in blue) and $\mathrm{ML}_{\mathrm{norm}}$ (in black) signals *(e)* Trajectory of the experiment on the SOM-2_b, where the black dots mark the projection of the discharge. Smaller dots indicate the beginning of the flat-top, and larger dots the final part of the discharge. The red dots mark the first transition in the disruptive region, while the yellow dot marks the last sample of the experiment projection.

addition to the mode-locking. Moreover, test samples falling into disruptive (red) clusters present low values of ML_{nm} ($ML_{norm} < 0.2$ mT/MA [33]), so that we can conclude that the SOM detects as disruptive other off-normal patterns in addition to the mode-locking, as for instance in the example in figure 7 between [13.75–13.88]s.

As a general comment, as also seen in [22], the absence of a peaking factor from the vertical camera makes the distinction between core and edge radiative phenomena ambiguous. In fact, there are cases where radiation blobs localized at the edge, not associated with increased disruption risk, occur in RTEs. In the bottom left region of the SOM-2, the presence of training samples from RTEs, even if they are very few, inhibits the alarm for DTEs. The presently adopted bolometer peaking factors only analyze the horizontal bolometer camera, and an upgrade of these signals would allow the vertical camera to unambiguously distinguish core and edge radiation. Future work will integrate the radiation peaking factors from the vertical camera of the bolometer, with the aim to better discriminate the core radiative phenomena from the edge ones.

Moreover, note that in the actual version of the algorithm the alarm criterion is simple, since the alarm is triggered when the sample enters the disruptive (red) region of the map, without implementing an assertion time or other complex techniques to enhance the SOM disruption prediction performance. Therefore, to enhance the performance of SOM, an analysis of cluster composition will be conducted, with the aim of labeling clusters based on statistical information criteria.

8. Conclusions

This work investigates the development of an unsupervised disruption predictor for JET, based on a SOM. The SOM model is able to map in an unsupervised way the high dimensional space of JET, in a 2D space. The map is then colored using only the information on the termination of the experiment, i.e. if it was regularly terminated or disrupted. Without assuming a priori knowledge on the appearance of disruption precursors, the SOM discovers non-trivial relationships and captures the complicated interplay of device diagnostics on the internal plasma states from the experimental data. The supplied model is highly interpretable: it is possible to visualize high-dimensional data and easily interrogate the model to

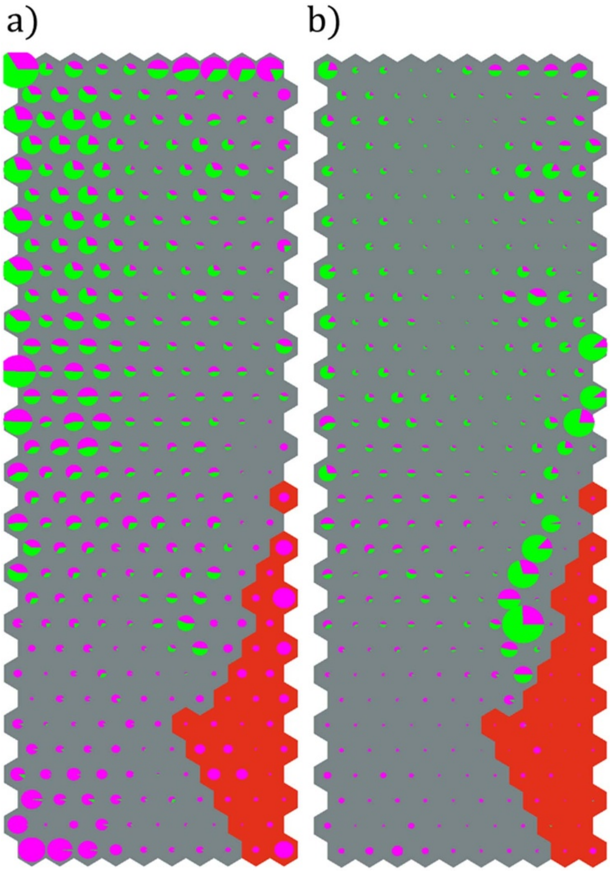

Figure 8. Pie-Chart representation of SOM-2_b of data distribution of different test sets: *(a)* training data; test data. *(b)* RTE samples are green; DTE samples are magenta.

understand why it makes the connections and correlations. In the SOM-2_b, it was possible to visualize where the samples from the different JET campaigns were projected, and the difference between the typical RTE and DTE evolutions. Two areas of interest were analyzed, suggesting that the additional information coming from the bolometer vertical camera may help discriminating core and edge radiative phenomena. Moreover, additional information could be used exploiting MHD spectrograms of the Mirnov coils, as in [5, 33].

Further work will be dedicated to the definition of equations describing the boundary of the different SOM regions, with the goal of defining interpretable boundaries between the regular and the disruptive terminations to be monitored during operation.

Future tokamak reactors, such as ITER, and existing tokamaks differ greatly in sizes, configurations, operation regimes, and plasma parameters values. In view of the application of data-driven disruption predictor to ITER, the interpretability of the model outputs is just as pivotal as achieving optimal predictor performance. Thus, it is important that present disruption predictors give insight into the instability mechanisms, to identify and explain disruption root causes and event chains in existing tokamaks. For its high interpretability, the SOM could be a valuable help toward this extrapolation to ITER, allowing to match operational parameters among tokamaks, scale physical laws, extract rules and identify common patterns in different devices.

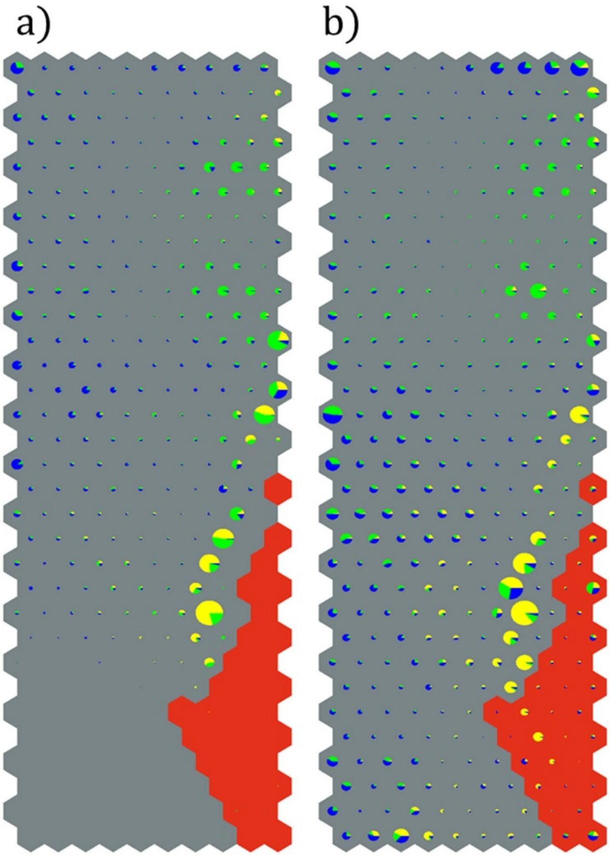

Figure 9. Pie-Chart representation of SOM_2b: *(a)* RTEs test data distribution during different experimental campaigns; *(b)* DTEs test data distribution during different experimental campaigns. The 2011–2013 campaigns are depicted in blue, the 2016 campaign is in green and the 2019–2020 campaigns are in yellow.

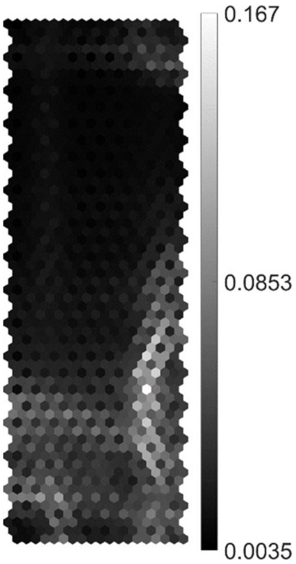

Figure 10. U-Matrix representation of the SOM-2_b.

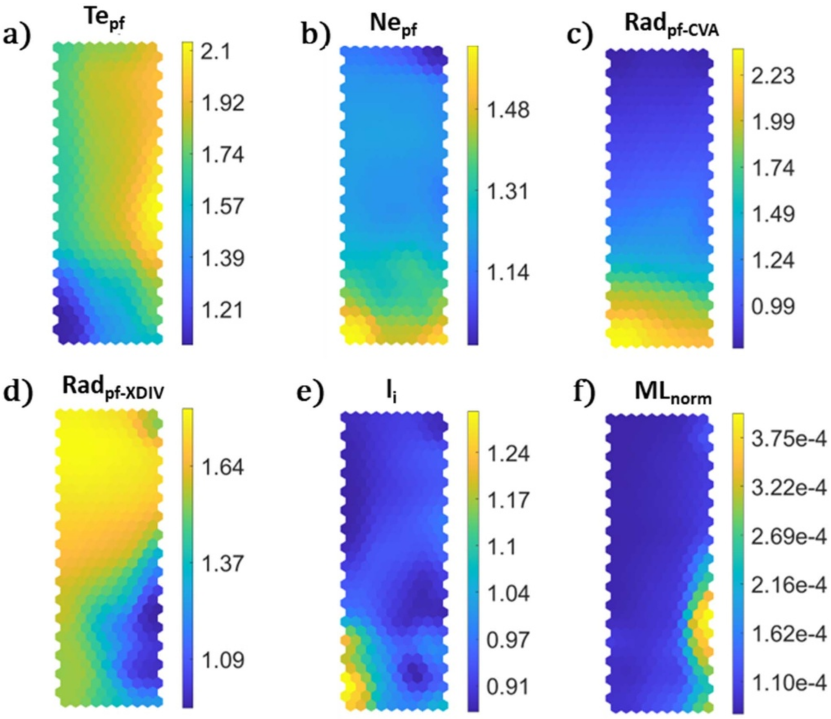

Figure 11. SOM-2. Component planes, indicating the distribution of the normalized input signal values: *(a)* Te$_{pf}$; *(b)* Ne$_{pf}$; *(c)* Te$_{pf\text{-}CVA}$; *(d)* Rad$_{pf\text{-}XDIV}$; *(e)* l$_i$; *(f)* ML$_{norm}$.

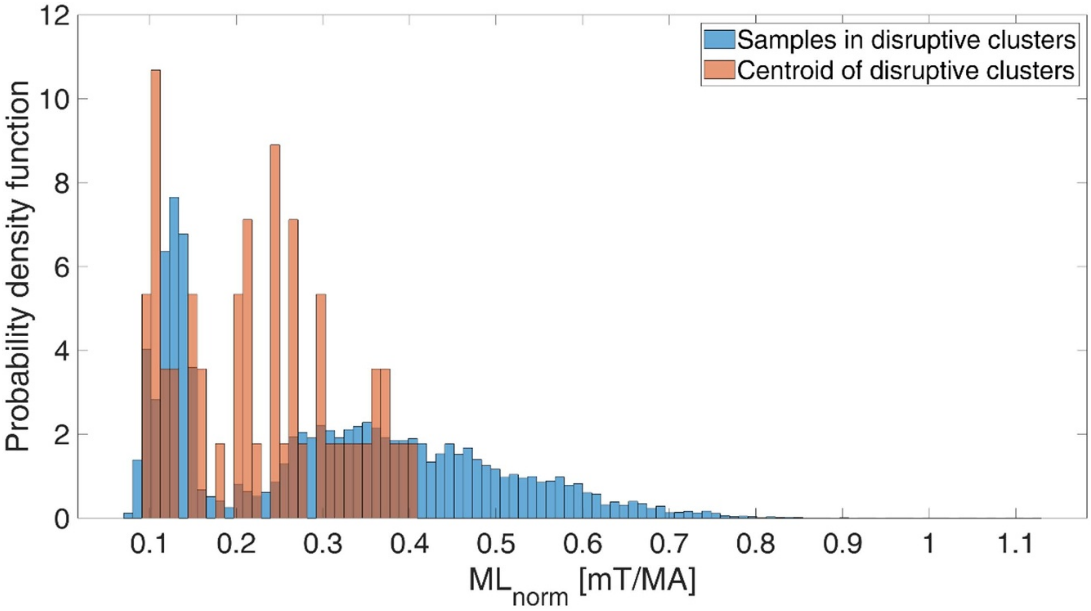

Figure 12. Probability density function of the ML$_{norm}$ values for the data samples and the centroids of the disruptive clusters.

Acknowledgments

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.

ORCID iDs

Enrico Aymerich https://orcid.org/0000-0003-3787-7685
Alessandra Fanni https://orcid.org/0000-0001-8604-5282
Fabio Pisano https://orcid.org/0000-0003-0162-0562
Giuliana Sias https://orcid.org/0000-0002-2289-301X
Barbara Cannas https://orcid.org/0000-0002-2766-0557

References

[1] Windsor C.G., Pautasso G., Tichmann C., Buttery R.J. and Hender T.C. (JET Contributors, ASDEX Upgrade Team) 2005 A cross-tokamak neural network disruption predictor for the JET and ASDEX Upgrade tokamaks *Nucl. Fusion* **45** 337–50

[2] Rattá G.A., Vega J. and Murari A. 2018 Viability assessment of a cross-tokamak AUG-JET disruption predictor *Fusion Sci. Technol.* **74** 13–22

[3] Zhu J.X., Rea C., Montes K., Granetz R.S., Sweeney R. and Tinguely R.A. 2020 Hybrid deep-learning architecture for general disruption prediction across multiple tokamaks *Nucl. Fusion* **61** 026007

[4] Kates-Harbeck J., Svyatkovskiy A. and Tang W. 2019 Predicting disruptive instabilities in controlled fusion plasmas through deep learning *Nature* **568** 7753

[5] Ferreira D.R., Martins T.A. and Rodrigues P. (Contributors JET) 2021 Explainable deep learning for the analysis of MHD spectrograms in nuclear fusion *Mach. Learn.: Sci. Technol.* **3** 015015

[6] Bonalumi L. *et al* 2024 eXplainable artificial intelligence applied to algorithms for disruptions prediction in tokamak devices *Front. Phys.* **12** 1359656

[7] Shen C. *et al* 2023 IDP-PGFE: an interpretable disruption predictor based on physics-guided feature extraction *Nucl. Fusion* **63** 046024

[8] Rea C., Montes K.J., Pau A., Granetz R.S. and Sauter O. 2020 Progress toward interpretable machine learning–based disruption predictors across tokamaks *Fusion Sci. Technol.* **76** 912–24

[9] Pau A., Fanni A., Carcangiu S., Cannas B., Sias G., Murari A. and Rimini F. 2019 A machine learning approach based on generative topographic mapping for disruption prevention and avoidance at JET *Nucl. Fusion* **59** 106017

[10] Pau A. *et al* 2018 A first analysis of JET plasma profile-based indicators for disruption prediction and avoidance *IEEE Trans. Plasma Sci.* **46** 2691–8

[11] Rattá G.A., Vega J., Murari A. and Gadariya D. (Contributors JET) 2021 PHAD: a phase-oriented disruption prediction strategy for avoidance, prevention, and mitigation in JET *Nucl. Fusion* **61** 116055

[12] Rossi R., Gelfusa M., Craciunescu T., Wyss I., Vega J. and Murari A. 2024 A hybrid physics/data-driven logic to detect, classify, and predict anomalies and disruptions in tokamak plasmas *Nucl. Fusion* **64** 046017

[13] Rea C., Granetz R.S., Montes K., Tinguely R.A., Eidietis N., Hanson J.M. and Sammuli B. 2018 Disruption prediction investigations using machine learning tools on DIII-D and Alcator C-Mod *Plasma Phys. Control. Fusion.* **60** 084004

[14] Cannas B., Fanni A., Pautasso G. and Sias G. 2011 Disruption prediction with adaptive neural networks for ASDEX Upgrade *Fusion Eng. Des.* **86** 1039–44

[15] Aymerich E., Fanni A., Sias G., Carcangiu S., Cannas B., Murari A. and Pau A. 2021 A statistical approach for the automatic identification of the start of the chain of events leading to the disruptions at JET *Nucl. Fusion* **61** 036013

[16] Aledda R., Cannas B., Fanni A., Pau A. and Sias G. 2015 Improvements in disruption prediction at ASDEX Upgrade *Fusion Eng. Des.* **96–97** 698–702

[17] Ai X.K. *et al* 2023 Tokamak plasma disruption precursor onset time study based on semi-supervised anomaly detection *Nucl. Eng. Technol.* **56** 1501–12

[18] Cannas B., Fanni A., Murari A., Pau A. and Sias G. (JET EFDA Contributors) 2013 Automatic disruption classification based on manifold learning for real-time applications on JET *Nucl. Fusion* **53** 093023

[19] Cannas B., Fanni A., Murari A., Pau A. and Sias G. (JET EFDA Contributors) 2014 Overview of manifold learning techniques for the investigation of disruptions on JET *Plasma Phys. Control. Fusion.* **56** 114005

[20] Cannas B., Fanni A., Murari A., Pau A. and Sias G. (The JET EFDA Contributors) 2013 Manifold learning to interpret JET high-dimensional operational space *Plasma Phys. Control. Fusion.* **55** 045006

[21] Aymerich E., Cannas B., Pisano F., Sias G., Sozzi C., Stuart C., Carvalho P. and Fanni A. 2023 Performance comparison of machine learning disruption predictors at JET *Appl. Sci.* **13** 2006

[22] Aymerich E., Sias G., Pisano F., Cannas B., Carcangiu S., Sozzi C., Stuart C., Carvalho P.J. and Fanni A. (The JET EFDA Contributors) 2022 Disruption prediction at JET through deep convolutional neural networks using spatiotemporal information from plasma profiles *Nucl. Fusion* **62** 066005

[23] Kohonen T. 1982 Self-organized formation of topologically correct feature maps *Biol. Cybern.* **43** 59–69

[24] Kohonen T. 1990 The self-organizing map *Proc. IEEE* **78** 1464–80

[25] Vesanto J., Himberg J., Alhoniemi E. and Parhankangas J. 2000 *SOM toolbox for Matlab 5* (Helsinki University of Technology, Finland) (available at: www.cis.hut.fi/somtoolbox/package/papers/techrep.pdf)

[26] Vesanto J. 1999 SOM-based data visualization methods *Intell. Data Anal.* **3** 111–26

[27] Ultsch A. and Siemon H.P. 1990 Kohonen's self-organizing feature maps for exploratory data analysis *Int. Neural Network Conf. (INNC-90), (Paris, France, 9–13 July 1990)* (Kluwer Academic) pp 305–8 (available at: https://archive.org/details/innc90parisinter0001inte/page/305/mode/2up)

[28] Kaski S. and Kohonen T. 1996 Exploratory data analysis by the self-organizing map: structures of welfare and poverty in the world *Proc. 3rd Int. Conf. on Neural Networks in the Capital Markets-Neuronal Networks in Financial Engineering* (*11–13 October 1995*), (World Scientific) pp 498–507

[29] Pucella G. *et al* 2021 Onset of tearing modes in plasma termination on JET: the role of temperature hollowing and edge cooling *Nucl. Fusion* **61** 046020

[30] Piron L. *et al* 2021 Progress in preparing real-time control schemes for Deuterium-Tritium operation in JET *Fusion Eng. Des.* **166** 112305

[31] Piron L. *et al* 2023 Radiation control in deuterium, tritium and deuterium-tritium JET baseline plasmas—part I *Fusion Eng. Des.* **193** 113634

[32] Sozzi C. *et al* 2021 Termination of discharges in high performance scenarios in JET *28th IAEA Fusion Energy Conf. (FEC 2020) (31 July 2024)* (available at: https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3320906)

[33] Aymerich E., Sias G., Atzeni S., Pisano F., Cannas B. and Fanni A. 2024 MHD spectrogram contribution to disruption prediction using convolutional neural networks *Fusion Eng. Des.* **204** 114472