IOP Publishing - International Atomic Energy Agency

Nuclear Fusion

PAPER

A machine learning approach based on generative topographic mapping for disruption prevention and avoidance at JET

To cite this article: A. Pau et al 2019 Nucl. Fusion 59 106017

Digital Object Identifier (DOI): http://dx.doi.org/10.1088/1741-4326/ab2ea9

A Machine Learning approach based on Generative topographic mapping for disruption prevention and avoidance at JET

A. Pau[1], A. Fanni[1], S. Carcangiu[1], B. Cannas[1], G. Sias[1], A. Murari[2], F. Rimini[3] and the JET Contributors*

[1] Electrical and Electronic Engineering Dept.-University of Cagliari, Piazza D'Armi, 09123, Cagliari, Italy.

[2] Consorzio RFX-Associazione - EURATOM ENEA per la Fusione, Padova, Italy

[3] CCFE, Culham Science Centre, OX14 3DB Abingdon, UK.

* See the author list of "X. Litaudon et al 2017, Nucl. Fusion 57 102001.

Abstract- Predictive capabilities better than 95%, and very limited false alarms, are demanding requirements for reliable disruption prediction systems in tokamaks such as JET or, in the near future, ITER. The prediction of an upcoming disruption has to be provided sufficiently in advance in order to apply effective disruption avoidance or mitigation actions preventing the machine to be damaged.

In this paper, following the typical machine learning workflow, a Generative Topographic Mapping (GTM) of the operational space of JET has been built using a set of disrupted and regularly terminated discharges. In order to build the predictive model, a suitable set of dimensionless, machine-independent, physics-based features have been synthesized, which make use of 1D plasma profiles information, rather than simple zero-D time series. The use of such predicting features, together with the power of the GTM in fitting the model to the data, allows obtaining, in an unsupervised way, a 2-dimensional map of the multi-dimensional parameter space of JET, where it is possible to identify a boundary separating the region free from disruption from the disruption region. In addition to helping in operational boundaries studies, the GTM map can also be used for disruption prediction exploiting the potentiality of the developed GTM toolbox to monitor the discharge dynamics. Following the trajectory of a discharge on the map throughout the different regions, an alarm is triggered depending on the disruption risk of these regions. The proposed approach to predict disruptions has been evaluated on a training and an independent test set, allowing to achieve very good performance with only one tardive detection and a limited number of false detections. The warning times are suitable for avoidance purposes and, more important, the detections are consistent with physics causes and mechanisms that destabilize the plasma leading to disruptions.

I. Introduction

Avoiding plasma disruptions will be one of the major concerns for the next generation of tokamaks such as ITER and DEMO. Disruptions, indeed, can cause severe damage to the structural integrity of the machines, forcing unexpected and eventually long maintenance interventions, which significantly reduce the availability of the device [1]. Avoiding disruptions or mitigating their effects requires quite different actions. For avoidance, the chain of events preceding the disruption has to be detected and suitable interventions have to be performed either to fully recover the nominal plasma parameters, keeping the plasma within the pre-programmed operation window, or at most to terminate it in a controlled way. To this purpose, sufficiently long warning times have to be provided to the disruption avoidance system. For mitigation, massive amount of gas or pellets can be injected into the plasma to increase the radiation, with the aim of reducing thermal and electromagnetic loads.

The plasma control system (PCS) on ITER, as well as the one on JET, will necessitate the availability of reliable triggers that have to satisfy different requirements, depending on whether they are intended for avoidance or mitigation purposes. The reliability and the effectiveness of such triggers can be defined in terms of warning time (that is the time between the trigger and the time of disruption), correct predictions, missed and false alarms. In the case of the disruption mitigation system (DMS), the requirements for the warning time are less stringent [2], and a mandatory limit is set by the system latency time. For the disruption mitigation valves (DMV) on JET, a working assumption for the minimum warning time is 10 ms, whereas, according to the present design, for the shattered pellet injectors (SPI) of the ITER DMS it will be 30 ms. In the case of

disruption avoidance, the minimum warning time depends on the combination of several factors, such as the type of disruption, the time scales involved by physics mechanisms playing a role, and the time required by the control actuators to intervene on the plasma state. As far as the success rate figures of merit are concerned, correct predictions and false alarm rates, again, are defined depending on the requirements of the specific machine. For ITER DMS, the requirements vary depending, on the one hand, on the machine operational phase and on plasma parameters and, on the other hand, on the disruption phenomena to be detected. During the operations at full performance, the failure rate in the detection of the current quench (CQ) and the vertical displacement event (VDE), which is needed to protect the machine against electromagnetic forces and the release of magnetic energy, should be less than 1%. The failure rate in the detection of the thermal quench (TQ), needed for the mitigation of the associated thermal loads, should be lower than 5%.

The disruption detection system in JET is presently based on the thresholds of single signals or a combination of them, such as the locked mode signal amplitude, the total plasma energy, the plasma current, the loop voltage signals, or the peaking of the radiated power. Nevertheless, these MHD indicators or plasma control parameters may not always be the best early predictors for the disruption onset, because of their late appearance in the chain of events leading to disruption. Hence the need to monitor other quantities, often closely linked to the physics mechanisms that destabilize the discharge, such as the main kinetic plasma profiles, the radiation distribution, and the internal inductance [3].

In addition to a real time protection and control system, an effective avoidance/mitigation system may foresee the presence of a real time disruption prediction block. The objective of the prediction block is essentially to recognize, in a reliable way, that is avoiding as much as possible false alarms, when the plasma is outside a stable and controllable parameter space. The corresponding operational boundaries, therefore, should identify the transition to a "disruptive parameter space", from which the plasma is expected to disrupt within a certain time. The new design of the Plasma Event and TRigger for Avoidance (PETRA), being implemented in JET, foresees, indeed, a cascade of a real time event detector block and the triggers handling to the already existent Real Time Protection System (RTPS) and Real Time Central Control (RTCC) units. The event detection system will comprise several disruption detection systems based on both physics-based disruption predictors and data driven predictors.

Up to now, disruption prediction systems for mitigation have been widely proposed on existing tokamaks. Physics basis for disruption prediction and detection have been discussed in [4],[5]. Several contributions have been proposed in the literature, aimed to develop disruption predictions using supervised data-based methods in JET [6], [7], ASDEX Upgrade [8], [9], J-TEXT [10], and DIII-D [11], only to quote some of them. Several plasma parameters, from a set of regularly terminated discharges, are used to describe the plasma operational space free from disruptions. Moreover, in order to describe the disrupted operational space, a set of disrupted discharges has been used and an *unstable pre-disrupted phase* has been statistically or heuristically identified and assumed equal for all the disruptions in the data base. This last choice introduces inconsistency in the prediction model that justify the prediction errors, even if they are generally limited to few tens of percent. These data-based models are unavoidably affected by the a priori selection of the training examples for the different classes they are supposed to predict. Indeed, the characterization of the boundary separating the different disruption classes might vary significantly, depending on the selection of the training examples; therefore, a proper selection of a pre-disruptive phase plays a key role on the performance of the model.

In addition, more general aspects concerning the generalization and the capability to extrapolate to new operational domains (new scenarios and/or new machines) have prevented machine learning methods from being considered as a viable and reliable solution for prediction and avoidance of disruptions. Firstly, as previously mentioned, they require an initial database to build and train the model. Secondly, all these data base models are strictly dependent on the choice of the parameters used to describe the plasma state. Therefore, they suffer the so-called "ageing" effect if they are used outside the operational domain of the training space. This parameter dependence affects these models as soon as the operational domain either evolves or changes significantly, for example moving from a machine to another. However, the consequent prediction performance deterioration can be limited, or ideally even avoided, provided that suitable signals are used to

describe the physics of the disruptions. There exists the possibility for an adaptive training almost from scratch [12] (that would be to some extent compatible with the gradual transition between the different operational phases foreseen for ITER), but the discussion of these methodologies is outside the scope of the present paper More recently, unsupervised manifold learning techniques, such as Self Organizing Map (SOM) and its probabilistic counterpart, the Generative Topographic Map (GTM), have been proposed to map the multidimensional plasma operational space in a reduced 2-dimensional space in JET [13] [14], and ASDEX Upgrade [15]. In particular, the GTM model has been used to classify disruptions in JET both with the carbon wall (CW) and the ITER like wall (ILW). Being an unsupervised method, to construct the map GTM does not theoretically require any assumption on the length of the pre-disruptive phase. In addition, belonging to the so-called generative models, GTM builds explicitly a density model defining probability distributions over the data and the manifold properties, providing at the same time a quantification of the uncertainty of the model fitted to the data.

In this paper, a disruption prediction system is presented, based on a GTM model. Particular care has been devoted to the selection and synthesis of the input plasma parameters, which aim to describe physics mechanisms characterizing disruptions. Leveraging parameters with a closer connection with such physics mechanisms can allow a more confident extrapolation to operational regions outside the training domain and to other tokamaks. The proposed approach is able not only to predict the disruptions, with warning times suitable for avoidance purposes, but, even more important, allows us to monitor the disruptions dynamics identifying often well in advance the causes of the discharge's destabilization, which are coherent with the physical phenomena leading to the disruptions.

II. Methodology

The next revolution, according to most experts, will come thanks to artificial intelligence (AI). Machine Learning (ML) is one of the disciplines of AI which "*gives computers the ability to learn without having been explicitly programmed*". It comprises a set of methods developed in recent decades in various scientific communities with different names such as: computational statistics, pattern recognition, artificial neural networks, adaptive filtering, theory of dynamic systems, image processing, data mining, adaptive algorithms.

The revival of interest in AI and machine learning is due to many factors, including the ever-increasing volumes and the variety of data available (the so-called Big Data), the reduction in the cost of computing resources and, at the same time, the exponential growth of computing power made available by cloud computing technologies and the storage of large amounts of data at affordable prices. All this allows to apply a variety of algorithms (already introduced since several decades) that implement complex mathematical calculations to large amounts of data for the resolution of real problems, even very complex.

Machine learning mainly concerns the use of algorithms for extracting knowledge from data. The choice to face a problem with machine learning depends on several factors: first of all, the problem must be complex, analytical models must not be available or they have to require too computationally demanding calculations, and it must involve multidimensional data, provided that these data are available in large quantities.

As it is well known, disruption prediction is a considerable complex problem. First principle models, which can reliably describe the phenomena leading to disruption with sufficient accuracy and with early enough warning, are not currently available. However, vast amounts of data are available that come from numerous diagnostics in years of experiments on different experimental devices. All the conditions for using ML techniques are therefore verified.

In Figure 1, the general machine learning workflow is reported. The first step in developing a ML application is the access, exploration and preparation of data that often come from different sources. Different types of data require different pre-processing techniques and, very important, a deep knowledge of the physics of the process that generated that data. Raw data must often be normalized, and outliers and offsets must be removed.

The subsequent phase consists of extraction and selection of features and/or transformation of data, to convert the raw data into information for the next phase of construction of the model. In addition, this phase has to be supported in a consistent way by a deep knowledge of the physics of the application domain. Furthermore, the features selection phase is of great importance, not only in terms of computational costs and data storage requirements, but especially to build simpler models that have less risk of overfitting and can reveal more clearly the underlying physics of the process.

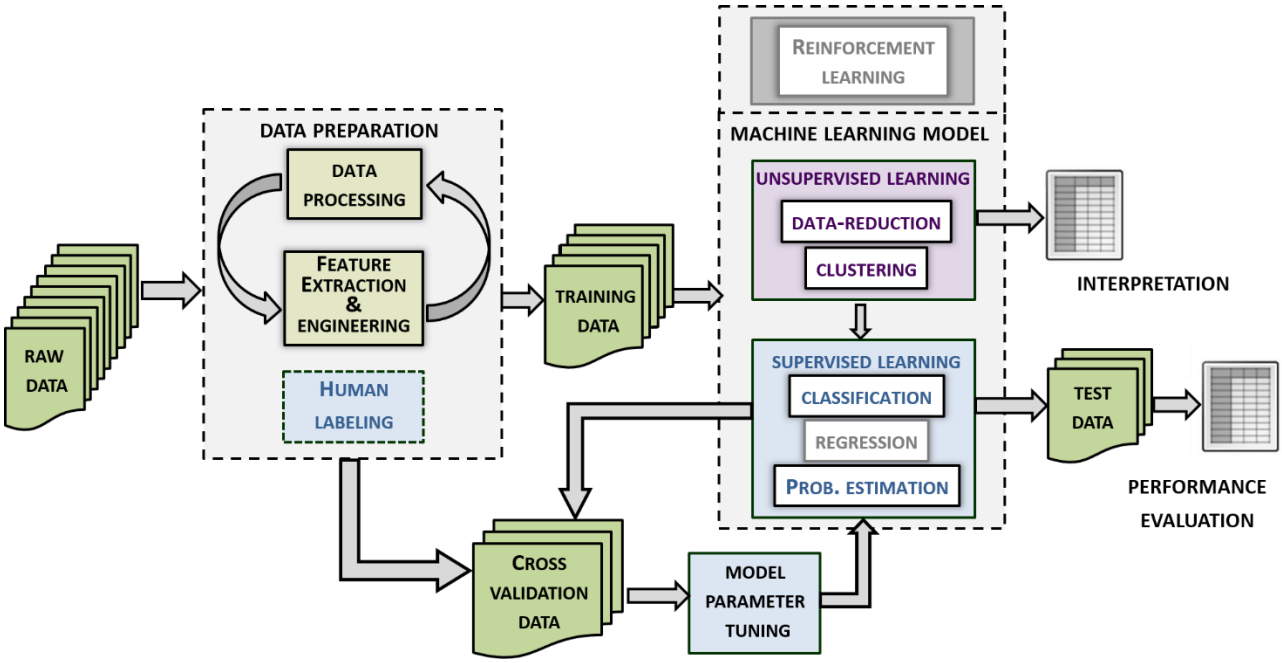

Figure 1 – Machine Learning workflow.

Once the data have been processed, the next phase concerns the development of the predictive model, which, among things, requires a proper identification of independent datasets to train and test such a model, as it will be described in section III. In this phase there is a great variety both in terms of available techniques and of parameters involved in these techniques. ML models can be roughly divided into supervised and unsupervised.

- In supervised learning, predictive models are obtained starting from the knowledge of both inputs and associated outputs, also called targets. These models can then make predictions on future outputs corresponding to inputs not used during the training phase of the model itself. Depending on the problem, for the development of the models, classification techniques can be used, which rank the data into categories, or regression models that predict continuous responses. If the output is a discrete value we have a classification problem, but if the output is continuous we refer to it as a regression problem. Among the supervised techniques, Support Vector Machines [16] are widely used for classification. Furthermore, regression algorithms such as logistic regression, decision trees, or neural networks are available.
- In the case of unsupervised learning, however, the task is to group the data using similarity among them. The input data set does not have any corresponding output values or labels and the objective, in this case, is to discover natural groupings and data patterns by investigating the structure of the input data. Clustering and data reduction techniques fall into the category of unsupervised techniques. Among them, k-Means methods, Self-Organizing Map [17] and its equivalent probabilistic version, Generative Topographic Map [18]. In this paper, the Generative Topographic Map has been employed, which belongs to the so-called Manifold Learning methods.

Potentially, a lot of information and features are available to build the disruption prediction model. Contrary to the typical operational diagrams, which are provided in terms of few plasma parameters to be easily visualized, the complex physics of disruptions takes place in high dimensional spaces. In order to deal with

such high dimensional spaces, the concept of manifold learning can be introduced: the data in high dimensional space can possibly lie in an embedded, eventually nonlinear, low dimensional manifold that can be easily visualized and understood if a two- or three-dimensional representation of the latent space is adopted.
Manifold learning algorithms [19] are indeed extremely helpful, because they are able to find a reduced dimensional space embedded in the high-dimensional data space, preserving topological and geometric properties of neighborhoods within data, and this has many implications from the point of view of computational efficiency, visualization, etc.. If requirements and objectives of disruption prediction are quite clearly defined, how to fully exploit the knowledge that can be extracted with machine learning and how to effectively connect it to the capability of the control system are still in an early stage of development. So far, the triggers provided by data-driven methods, as discussed in the introduction, have been developed mainly to satisfy the basic requirements of reliability and minimum warning time for mitigation purposes, without any ambition to saying anything about causes or main mechanisms developing in the pre-disruptive phase. This is something mandatory as a last layer of defense to protect the integrity of the machine, but for avoidance purposes some more knowledge about the nature of the trigger would be highly desirable. In fact, in order to be able to recover a discharge when a disruptive behavior has been recognized, the control system needs to know possibly cause and nature of the "off-normal" event(s) to react accordingly. This is the reason why recognizing the type of disruption and tracking the corresponding chain of events would be extremely important for avoidance purposes. This is also the reason why, from a more technical machine learning point of view, we are interested not only in a pure classification task (that can be done in general with any discriminative model, even if with different performance) but also in the properties of the parameter space where the relevant disruption physics takes place, its visualization and interpretative analysis. Manifold learning algorithms are intrinsically oriented to this type of analysis and, among those available, GTM seems to be the ideal candidate to the problem at hand because of the reasons that are discussed below.
Of course, GTM is not the only algorithm capable to do such analyses [20] but, taking into account also other aspects such as the computational efficiency, the possibility to deal with large datasets, etc., a general-purpose tool based on the GTM algorithm has been developed for analysis and visualization purposes [21], and is being further developed adding new features for a more advanced investigation of the mapped parameter space. In the following subsection, the fundamental concepts of the GTM are reported.
The last step of the machine learning process is the evaluation of the performance. New examples, not used to build the prediction model, are tested on the trained model and its generalization performance is estimated by using some performance indexes.
In the rest of the paper, the customization of the different steps of the machine learning workflow will be described, referring to the development of a reliable disruption predictor to be integrated in the real time control system of JET. An interdisciplinary approach has been adopted where machine learning and statistical tools have been supported with a deep physics knowledge of the disruption phenomena.
In this paper only JET data has been considered even if the general proposed scheme could be easily customized to different tokamaks, such as AUG or TCV, with particular reference to the extrapolation to next step fusion devices such as ITER and DEMO.

Generative Topographic Mapping

GTM is an advanced manifold learning algorithm that is able to compute, in an unsupervised way, a mapping from a low dimensional latent space into the high dimensional data space, preserving the topology of this latter. This means, on the other hand, that points close to each other in the data space will be mapped still close in the latent space. As already mentioned in the introduction, it is a generative model, which means it defines probability distributions over the data or over the manifold properties; it provides measures of uncertainty on the manifold and on the locations of the embedded points. Moreover, GTM fits into the framework of probabilistic theory and statistics, or in other words, it provides the possibility to exploit well-founded theory for fitting models to data, combining models, treatment of incomplete data, etc.. Conversely, algorithms like

Locally Linear Embedding (LLE) [22] try to preserve local linear relationship, minimizing the distortion of local derivatives. Algorithms like Multidimensional Scaling (MDS) or Isomap [23] (which extends MDS by incorporating the geodesic distances imposed by a weighted graph) perform low-dimensional embedding based on the pairwise distance between data points, so, basically, they try to preserve local distances.
Let $\boldsymbol{X} = \{\boldsymbol{x}_1;\ \boldsymbol{x}_2; \cdots \boldsymbol{x}_K\} \in \Re^L$ be a regular grid of nodes in the latent space, and $\boldsymbol{T} = \{\boldsymbol{t}_1;\ \boldsymbol{t}_2; \cdots \boldsymbol{t}_N\} \in \Re^D$ be the training data set in the data space, the basic idea behind GTM algorithm is to achieve a nonlinear mapping from $\boldsymbol{X}$ to $\boldsymbol{T}$ by using a linear combination of M Radial Basis Functions (RBF) Φ:

$$\boldsymbol{t}(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W} \cdot \Phi(\boldsymbol{x})$$

As suggested in [18], the basis functions are typically assumed to be radially symmetric Gaussians, whose centers are distributed on the uniform grid in $\boldsymbol{x}$-space, with a common width parameter σ, which determines the smoothness of the manifold.
In order to take into account the probability that actual data points belong to the latent space, a Gaussian noise is added to the data points $\boldsymbol{t}$ rendering the manifold a mixture of Gaussians:

$$p(\boldsymbol{t}|\boldsymbol{W}, \beta) = \frac{1}{K}\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} p(\boldsymbol{t}|\boldsymbol{x_k}, \boldsymbol{W}, \beta)$$

where β is the inverse of the noise variance. The adaptive parameters of the model ($\boldsymbol{W}$ and β) can be calculated by maximizing the log likelihood function by means of the Expectation Maximization (EM) algorithm [24]:

$$\max_{\boldsymbol{W},\beta} l = \sum_{n=1}^{N} ln\left(\frac{1}{K}\sum_{k=1}^{K} p(\boldsymbol{t}|\boldsymbol{x_k}, \boldsymbol{W}, \beta)\right)$$

In order to visualize the whole data in the map, the posterior probability distribution over the latent space is usually summarized through a statistical measure such as the mean or the mode:

$$\boldsymbol{x}_n^{mean} = \sum_{k=1}^{K} \boldsymbol{x}_k \cdot p(\boldsymbol{x}_k|\boldsymbol{t_n})$$

$$\boldsymbol{x}_n^{mode} = \sum_{k=1}^{K} argmax\ p(\boldsymbol{x}_k|\boldsymbol{t_n})$$

where

$$p(\boldsymbol{x}_k|\boldsymbol{t_n}) = \frac{p(\boldsymbol{t}_n|\boldsymbol{x}_k, \boldsymbol{W}^*, \beta^*) \cdot p(x_k)}{\sum_{k'=1}^{K} p(\boldsymbol{t}_n|\boldsymbol{x}_k, \boldsymbol{W}^*, \beta^*) \cdot p(x_{k'})}$$

After learning the model, the posterior probability of a class, given the latent space, can be used to estimate the class probability of a new point in the test set. The class with the largest probability is assigned to the test point.

III. Data Base

The first step of the ML approach is the construction of a reliable and representative database to be used to build the prediction model. This is the most important and difficult phase because, despite the huge amount of data coming from several diagnostics, some of them can be redundant or even useless to describe and discriminate a disruptive behavior.

In the case of disruptions and, more in general for fast transient events, a standardized definition of characteristic times univocally defining phases of interest is a fundamental requirement to make analyses consistent across different devices. Furthermore, since transient events like disruptions inherently involve a large change of plasma parameters on a very short time scale, measurements and calculations accuracy plays a key role and can significantly affect analyses. Another important aspect, assuming the technical feasibility of the construction of a database with predefined characteristics, is the unavoidable presence of errors or potential inconsistencies, even after a very time-consuming manual analysis.
In order to deal with such critical aspects, a tool (*DIS_tool* [25]) able to process and correlate the measurements of several diagnostics for the detection of fast transients has been designed. In particular, *DIS_tool* synthesizes the complexity of the disruptive process implementing coherent definitions of characteristic times and parameters such as the thermal quench (T_{TQ}), the current quench time (T_{CQ}), the time of disruption (t_D) and the Mode Lock time (T_{LM}), which is the time where the locked mode amplitude starts to rise.
In this paper, by making use of *DIS_tool*, corroborated by a manual analysis, the main precursor phases of the disrupted discharges have been examined in detail, determining, among the others, also the time (T_i) that corresponds to the start of the chain of events leading to disruption. In the following, we refer to it as the *Reference Warning Time*.
Data for this study have been selected from the ITER Like Wall (ILW) experimental campaigns performed at JET from 2011 to 2013. In particular, 132 disrupted terminations and 114 non-disrupted terminations have been selected. These are mainly flat top disruptions, excluding those terminated by massive gas injection. It is worth mentioning that in the first JET-ILW campaigns, the use of the DMV was not mandatory, therefore most of the disruptions had the possibility to evolve naturally until the final loss of the plasma current. If on the one hand such a selection does not include more recent high-power campaigns, on the other hand it safeguards us from introducing any bias related to discharges terminated prematurely by massive gas injection, as compulsory for more high-performance plasmas. Taking into account the natural learning curve for the exploration of more advanced operation with the ILW, due to initial (probably) conservative assumptions about triggers and corresponding thresholds for the MGI activation, there is the possibility that part of the discharges was terminated by MGI when not yet absolutely necessary. In this first stage, from the point of view of the interpretative analysis as well as for the error analysis, it is extremely important to have a clear picture of the data involved in the analysis.

Diagnostics and feature engineering

In order to effectively extract the information associated with multi-dimensional signal data, one-dimensional profiles describing the evolution in time of basic plasma quantities such as the electron temperature, the electron density and the radiation have been processed synthesizing physics-based indicators to be provided as input features to the disruption predictor. In particular, as described in [3], the so called "peaking factors" of temperature (Te_{pf}), density (Ne_{pf}) and radiation (Rad_{pf}) have been computed and statistically analyzed, showing a high discrimination capability. The majority of the disruption predictors, proposed in the past, mainly rely on zero-dimensional MHD markers related to still rotating modes and, especially, to locked modes, which are basically the final precursor of most of the disruptions. Nevertheless, in many cases, the warning time is still unsatisfactory with respect to avoidance requirements, and a significant improvement can be reached using a more structured information as predictor inputs, such as the spatial distribution of kinetic quantities, current profiles and radiation.
Concerning the profile of the electron temperature, both the ECE (Electron Cyclotron Emission) and the High-Resolution Thomson Scattering diagnostics [26] satisfy basic requirements to allow, in principle, the calculation of comparable peaking factors. JET ECE heterodyne radiometer consists of 96 channels over 4 data acquisition bands, allowing either first harmonic measurements (O-mode) or second harmonic measurements (X-mode). The High-Resolution Thomson Scattering (HRTS) diagnostic on JET measures electron temperature (Te) and electron density (Ne), providing 63 data points per profile with a repetition rate of 20 light pulses per second (20Hz). The spatial resolution of the measurements for the core region and the

pedestal is respectively of 1.6 cm and 1 cm. Note that, HRTS has a lower time resolution than the ECE signals. However, T_e peaking factors based on the ECE diagnostic, in a not negligible number of cases, were found to be affected by cut-off of several channels. The effect was sometimes marginal, other times was heavily jeopardizing the calculation of the peaking factor itself. Therefore, it was decided to replace ECE measurements with those of the HRTS, that in next JET campaigns is supposed to be available in real-time. In this first stage, it is indeed much more valuable to get correctly the overall picture of the parameter space where the relevant physics takes place rather than restricting the analysis with respect to real-time requirements. The same considerations apply also to other parameters that are introduced in the following, such as the safety factor on magnetic axis, for which the post-processed signals have been used.

Concerning the radiated power, the main-vessel bolometric camera with a horizontal view of the plasma cross-section (KB5H) has been used. The camera collects the radiation along 24 chords, 8 of which in each case cross the divertor region and the region adjacent to the divertor with 8 cm separation between the chord axes. The other 16 channels cover the entire plasma. A simple pinhole structure is used to define the lines-of-sight of the camera [27].

In [3], the peaking factors have been considered as features defined as a "*core versus all*" metric, i.e., they are defined as the ratio between the mean value of the considered radial profile (temperature, radiation, density) around the magnetic axis and the mean value of the measurements over the entire radius. The radial interval to define the "core" has been empirically set equal to 25% of the radial axis (the minor radius for the horizontal mid-plane temperature and density measurements and the vertical semi-axis of the poloidal cross section for the horizontal bolometer camera radiation profile).

Regarding the peaking factor of the radiation, with respect to the initial unique parameter [3], two different indicators have been derived, splitting the information carried out by the global radiation distribution according to the two main mechanisms involving a radiation collapse: the accumulation of high-Z impurities in the plasma core as opposed to edge-radiative collapse. As analyzed in [3], the main mechanisms with which the discharge is being destabilized are quite different, as well as the time scales involved in the corresponding chain of events. Nevertheless, although the initial version of the peaking factor was such that both the edge and core radiation collapse could be detected quite nicely, the two mechanisms are not mutually exclusive, and even if developing on different time scales, there are cases where both are simultaneously affecting the discharge.

As a general comment, even though in many cases it is possible to find "clean" examples of a well-defined chain of events corresponding to a specific disruption type, in other cases there is an interplay of more than one mechanism destabilizing the discharge, so that synthesizing more detailed and targeted indicators goes in the direction of a more accurate and flexible avoidance and prediction system.

Therefore, the information carried out by the peaking factor of the radiation profile is decomposed in two separated indicators, one always based on the "core vs all metric" ($Rad_{pf\text{-}CVA}$) but having subtracted the radiation in the X-point/divertor region, and the other one based on the "divertor vs all" ($Rad_{pf\text{-}XDIV}$) metric but having subtracted the radiation in the core (decoupling in this way the contribution of the core radiation from the contribution of the divertor radiation). This has allowed in a sense, a decoupling of the two behaviors, improving at the same time the resolution of each of them.

To the five 1-D profile indicators (including the internal inductance as representative of the current density profile), other two dimensionless parameters have been integrated in the final dataset: the fraction of radiated power with respect to the total input power (P_{frac}) and the safety factor on magnetic axis (q_{AX}). The first one is a well-known indicator of the power balance and of the global radiative collapse, and, together with the two aforementioned radiation peaking factors "locally" defined, has the function to connect a spatial information (related to the 1-D profile of the radiation) with the entity of the radiation collapse in macroscopic terms. The second one is an important equilibrium parameter and, as well-known from MHD stability theory, it is connected to the presence of the resonant surface for q=1 and the sawtooth crashes due to the instability of the internal kink mode (m=1, n=1). This information, in connection with the peakedness of the current profiles

given by the internal inductance (L_i), and the other parameters plays a key role on the plasma stability as it will be described with some examples in the next sections.
As it will be discussed in the conclusions, some of these quantities might be affected by not negligible uncertainties in real-time processing. Nevertheless, the main objective of this analysis is to show the possibility to identify and exploit operational boundaries in a reduced set of physics-based dimensionless parameters for disruption avoidance. This preliminary step is needed to assess the feasibility of the approach that, at this first stage, has to be investigated separately from the uncertainties in real-time measurements.
The GTM models, as it will be described in the following sections, have been built using the flat top phase of non-disrupted discharges and the unstable phase of the disrupted discharges. This last phase has been defined as that after the time T_i, which has been identified as the start of relevant chain of events that is destabilizing the discharge, and up to the disruption time t_D. For the considered discharges, these stable and unstable phases have been uniformly sampled and the obtained samples have been labelled as Safe Samples (SS), and disruptive samples (DS), respectively.

Feature preprocessing

The selection of the discharges included in the database, both disruptions and regular terminations, has been carried out trying to preserve statistically the overall variety and variability of scenarios and experiments that were carried out during the considered experimental campaigns. The main constraint, as described in [3], was represented by the availability and the consistency of the signals needed to compute the features described in the previous section.
After the shot selection, the signals have been resampled with a uniform time step of 2 ms, which corresponds to the cycle time of the JET ATM network for real-time control. The basic preprocessing that has been performed takes into account real-time requirements, so that possible spikes and outliers are either discarded or smoothed out, by a causal median filtering of 40 ms width applied to each of the features. In fact, since the main objective of this work is to develop a tool for disruption avoidance, we are not interested in fast transient phenomena, but rather in destabilizing mechanisms that change the plasma state over longer time scales. A reasonable choice of the preprocessing parameters is a key element of an avoidance/prediction system to guarantee reliable detections and to avoid false alarms.
Then, a subsequent step, which is part of almost any machine learning workflow, consists of reducing the amount of data to be computationally manageable and also to overcome the classes unbalance that is due to longer stable phases of non-disrupted discharges with respect to the unstable phases of the disrupted ones. As previously cited, these phases have been uniformly resampled, so the non-disrupted space would be over represented with respect to the disrupted space.

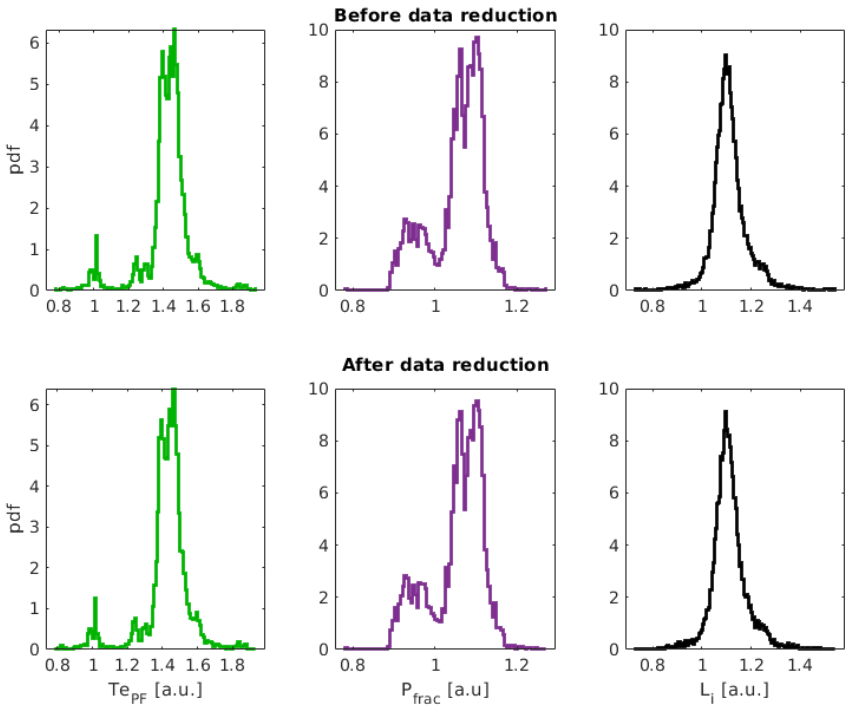

Figure 2 – Probability Density Functions of Te_{pf}, P_{frac}, and L_i, in arbitrary units [a.u.], before and after the data reduction for the non-disruptive space: the statistical distribution is preserved.

A GTM-based data reduction algorithm has been then applied shot by shot to the non-disrupted discharges reducing the number of samples by a prefixed factor. The data reduction algorithm allows to reproduce the same probability density function of the original data, as shown in Figure 2 that reports the probability density function of some of the input parameters of the original data set, on the top, and of the reduced set, on the bottom, using a reducing factor of nine. The same behavior has been found for all the other considered parameters. The peculiarity of preserving the probability distributions of high dimensional data is extremely important to avoid either the loss or the partial distortion of the statistical properties of the initial population and is not necessarily achieved by random under sampling. The algorithm, iterating over all the shots included in the training dataset, performs a 2D mapping producing an unsupervised clustering of the data, with hyper-parameters automatically selected on the base of the shot data structure. For each cluster of a "shot-map", the procedure retains a number of samples reduced by the data-reduction factor (that is 9 in our case). The reduced number of samples is selected with respect to their ordered rank distances from the barycenter of the cluster itself: samples regularly spaced in such a rank allows us to preserve both the variability of data in the cluster and the probability density distribution of the input features. These two aspects play a key role in the effectiveness and the validity of generative models.

IV. Analysis of JET operational space with GTM

All the synthesized features have been used as input to build the GTM model. In order to test the generalization capability of the GTM model as disruption predictor, the data base has been split into two sets, a training set, which has been used to build the map, and a test set, to evaluate the prediction performance. The training set contains 89 disrupted shots and 70 regular terminations, whereas the test set is composed by 43 disruptions and 44 safe shots, which have never been presented to the model during the training. The selection of the training and test set has been performed in order to obtain two sets maximally independent and representative of the operational space. To this purpose, the test set contains discharges temporally subsequent to those in the training set. The different disruption classes are represented with the same percentage in the two sets, and repetition of discharges with same or similar setting are avoided.

A multi-objective Tabu Search (TS) [28] procedure has been customized to optimize the free parameters of the GTM, i.e., the map dimension *K*, the number of radial basis functions *M*, and their width σ. The aim of the optimization process is to maximize the log likelihood of the mapping of the training set, minimizing at the same time its sparsity, defined in terms of the percentage of empty clusters.
The TS is a metaheuristic method that looks for the optimal solution of an optimization problem by exploring the search space based on the use of adaptive memory. Starting from a random or a given initial point (described by the three free parameters of the GTM), at each iteration the algorithm explores all the neighbor points, selecting the best one. The TS strategy consists in exploring the search space along the coordinate directions (in this case each coordinate corresponds to a parameter of the GTM) and taking into account the most promising points to follow the search. TS avoids cycling by keeping in memory a list of examined points or their features so that the exploration of already investigated regions of the search space is inhibited and stored in a Tabu List. The resulting optimal GTM has *K*=2500 latent points, and *M*=400 radial basis functions with $\sigma = 0.8$, which correspond to a log likelihood of $9.85 \cdot 10^5$ and a percentage of empty clusters of 12%, obtained after 30 TS iterations.
Figure 3 reports the 2-D GTM of the 7-D JET operational space corresponding to the feature set described in the previous section. Figure 3-a reports the pie-plane representation, with reference to the frame highlighted in Figure 3-b, of the mode of the posterior probability distribution over the latent space. Each node, or cluster, in the map represents the pie chart of the samples associated to that node. Green color refers to samples belonging to non-disruptive discharges, red color refers to samples coming from the unstable phase of the disrupted discharges. In Figure 3-b, each node in the map is colored depending on its composition: the green clusters contain only samples, the red clusters contain only samples (whereas grey clusters contain both non-disruptive and disruptive samples. The white clusters are empty. As can be seen, a well-defined separation between the two regions representing the disruptive (red) and non-disruptive (green) classes can be recognized in the 2-D latent space.

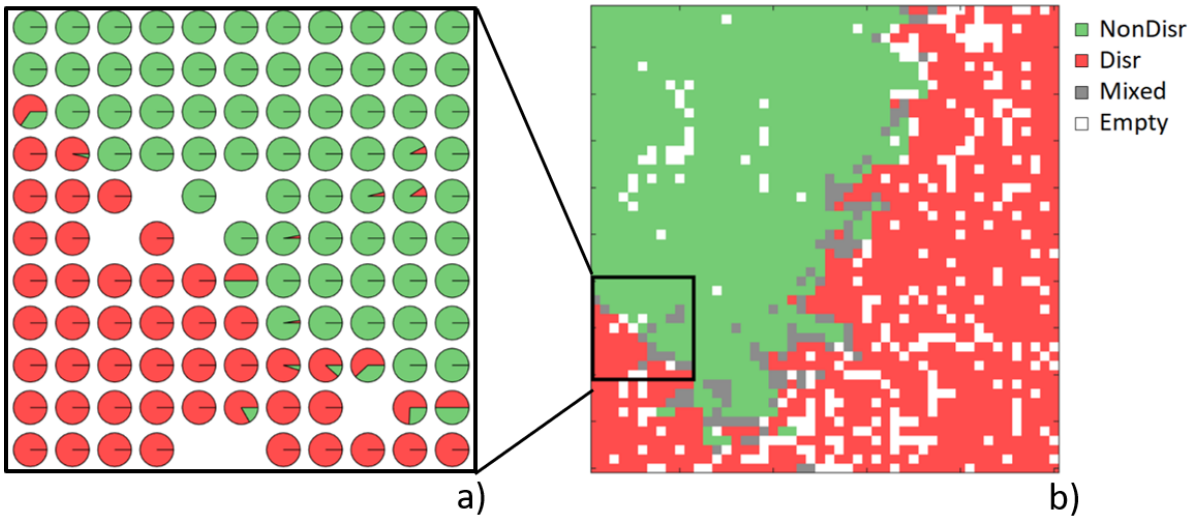

Figure 3 – GTM of the 7 plasma dimensionless parameters: (a) Pie plane representation (zoom of the frame selected in (b)) and (b) map colored on the basis of the node composition (each latent point is associated with a node).

It can be seen how the operational spaces of the regular terminations and disruptions are well confined and quite compact; besides some isolated spots, there is only a narrow overlap on the boundary separating the two classes. This great separability of the two regions suggests the possibility to exploit with high-performance the obtained 2-D map as disruption predictor, as will be presented in the following Section V.
A further figure of merit to evaluate locally the degree of separation between classes is to analyze the composition of the clusters in the map. As can be seen by analyzing the map composition reported in Table 1, the degree of "separability" of disrupted and non-disrupted samples is quite high, the percentage of empty

clusters is less than 12%, whereas the percentage of mixed clusters is less than 5%. The disrupted and non-disrupted clusters tend to aggregate according to the self-organization of the map that is driven only by similarities and differences in the probability density distribution of the input features. The unsupervised learning is shaping the 7-D input space in such a way that each of the two classes results to be predominant with respect to the other in different regions of the 2-D latent space. The optimal map has also a quite low sparsity.

Table 1. GTM composition

Type of cluster	# clusters	% clusters	% samples in the clusters	% samples of a certain class belonging to clusters of the same class
Disrupted	1053	42.12	47.96	95.92
Non-disrupted	1045	41.80	47.21	94.42
Mixed	107	4.28	4.83	-
Empty	295	11.8	-	-

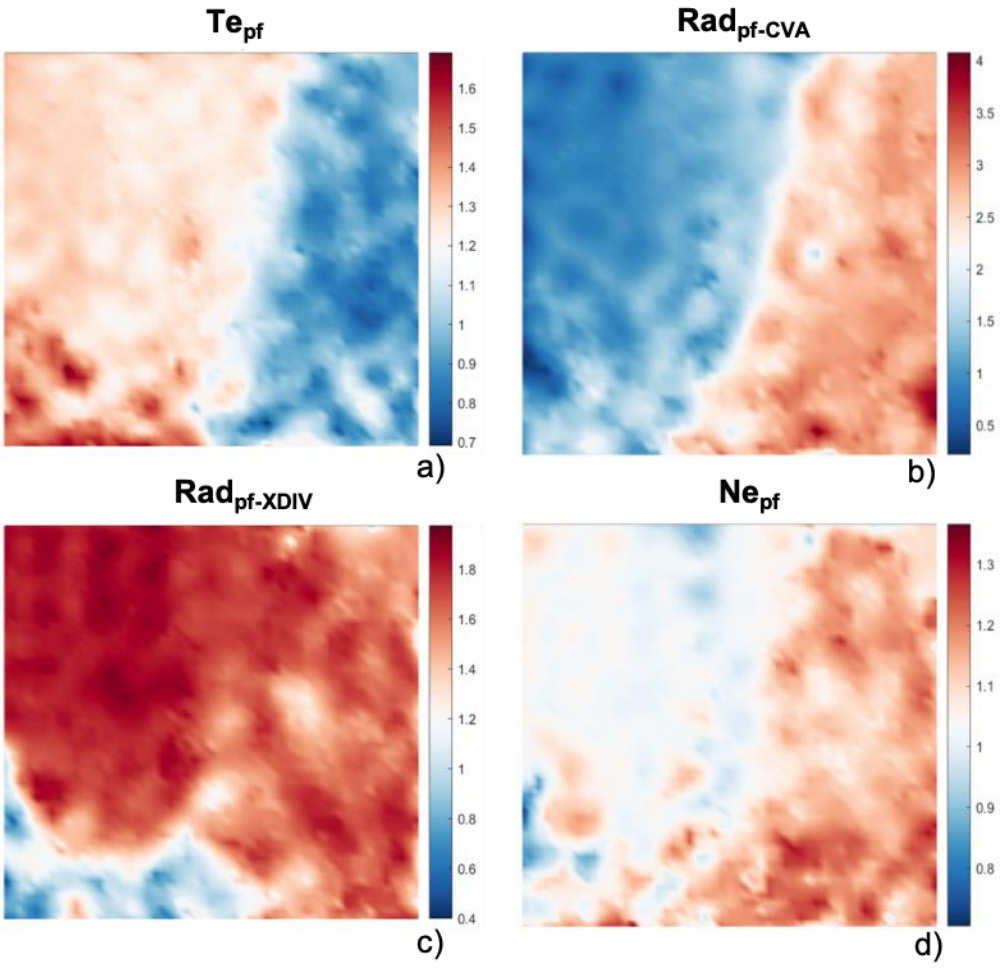

Figure 4 – Component plane representation of: (a) Peaking factor of the temperature (Te_{pf}); (b) Peaking factor of the radiation ($Rad_{pf\text{-}CVA}$); (c) Peaking factor of the radiation ($Rad_{pf\text{-}XDIV}$); (d) Peaking factor of the density (Ne_{pf}).

Figure 4 and Figure 5 report the Component Planes that represent the relative component distributions of each of the input parameters used for the mapping. Component Planes distributions reflect univariate probability distribution (pdf) information uncovering patterns in the data. As discussed in [3], it can be seen that the

individual features are already very descriptive of the different disruptive and non-disruptive behavior but what really makes the difference in the discrimination between the two spaces is how these parameters combine to capture the mechanisms destabilizing the discharge.

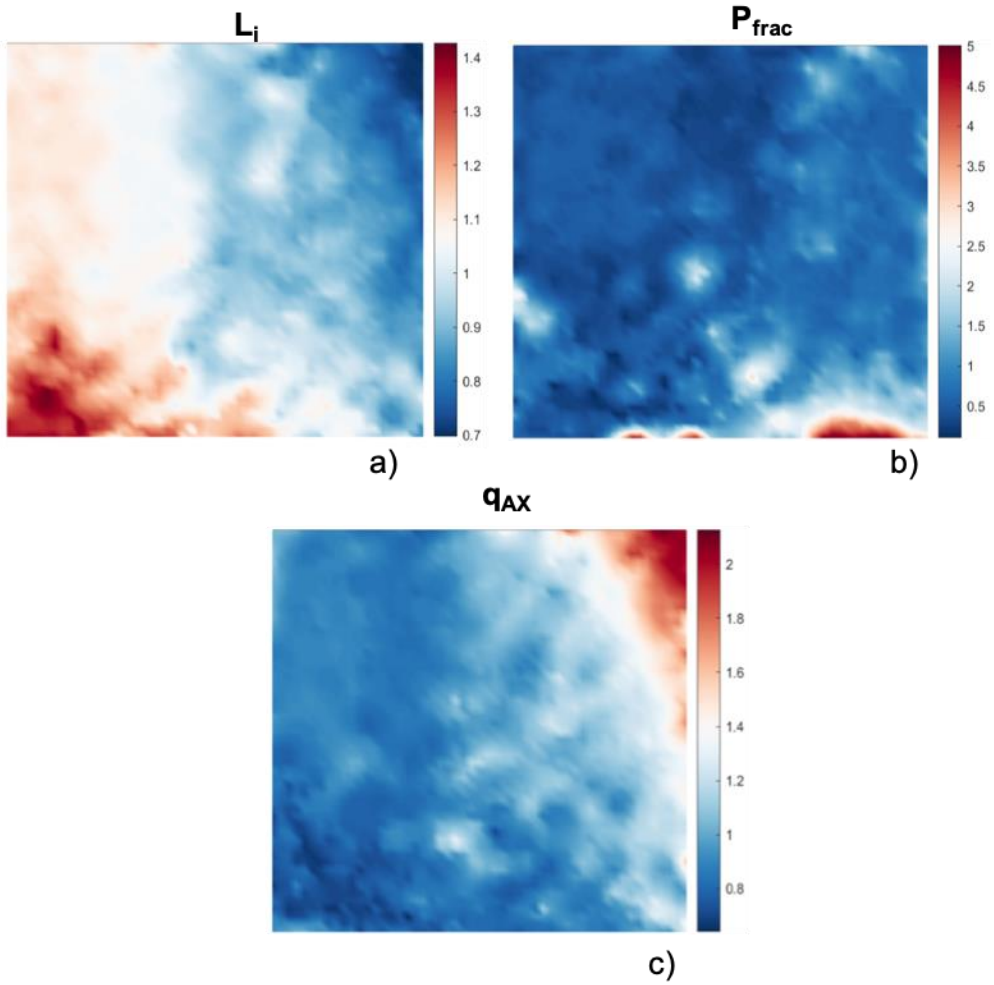

Figure 5 – Component plane representation of: (a) Internal Inductance (L_i); (b) ratio of the total radiated power and the total input power (P_{frac}); (c) safety factor at the axis (q_{AX}).

The basic interplay of radiation, kinetic and current density profiles have been already discussed in [3], whereas the contribution of the other components has been already described in the section dedicated to the feature construction. How each of the individual components resemble the shape highlighted in Figure 3 for the two classes depends on different factors. For instance, more than half of the disruptions composing the database are due to impurity accumulation and radiative collapse in the plasma core [13], [29], so that the distribution of Rad_{pf}-$_{CVA}$ reflects quite closely the disruptive region for high values of this quantity. In this case, the destabilizing mechanism is well described by such transition to higher values but is not the only ingredient and a fixed threshold is not enough for a reliable detection. The plasma response depends case by case by the plasma underlying conditions, the possible presence of other factors preventing, for instance, the full development of hollow temperature and current profiles. As can be seen by analyzing the component planes with respect to the regions occupied by the disruptive and non-disruptive samples, the patterns defined by the mapping cannot be easily extrapolated by considering individually the different features. The additional value of this machine learning approach is the capability to handle intrinsically the multivariate nature of complex operational spaces. This is extremely helpful for statistical analysis and visualization of the most recurrent patterns reflecting physics underlying mechanisms as well as for the interpretation of complex dependencies and relations among the different features.

As previously mention, the GTM is an unsupervised generative model, which aggregates the data, based on their probability density distributions. The map coloring used in this paper aims to highlight the discrimination capability of the map in terms of classification between disrupted and non-disrupted space. Moreover, as shown in Figure 6, which is a different visualization of the GTM of Figure 3, the same boundary between disrupted and non-disrupted spaces is clearly visible both on the mode representation of the GTM in Figure 6 a), and in the U-Matrix representation in Figure 6 b). In the GTM tool [21] used in this paper, the like-Unified Distance-Matrix (U-Matrix) plots the mean Euclidean distance between the barycenter of the samples of each cluster to its neighbors in a grayscale image. Macro regions of plasma states characterized by limited local variations are represented by lighter clusters and demarcated by darker clusters, which correspond instead to steeper local changes in the parameter space. This is consistent with the interpretation of the evolution of the plasma state significantly "deviating" during the transitions from a stable to an unstable phase. This result is particularly relevant since it has been obtained without any assumption a priori about the different nature of the two classes: it does represent an intrinsic property of the manifold where the data naturally lie. The correspondence of such intrinsic boundary separating the two regions corroborates the assumptions made with the manual selection of the Reference Warning Times (T_i).

Independently on the consistency in the selection of the warning times, one can wonder if a so clear separation and the resulting boundary are due to the global difference between non-disruptive and disruptive discharges rather than between the stable and the unstable phases. Therefore, in order to check how reliable the obtained boundary is, the non-disruptive discharges have been replaced by the stable phase of disruptions. Figure 7 a) reports the mode representation of the GTM obtained with only the disruptive discharges, whereas Figure 7 b) reports the corresponding U-Matrix. Note that, in this case, the obtained mapping is just representing the operational space of the disruptions.

As expected, the boundary is not as sharp as in the previous case and there is a larger overlapping because of the smoother transition between stable and unstable states but what is important to highlight here is that the boundary has still a similar shape. This is a clear indication that there is a common underlying structure characterizing the stable phases in the two cases. This is not obvious, and whether the "stable" phase of disruptive discharges is really close to that of non-disruptive discharges is still a debated point.

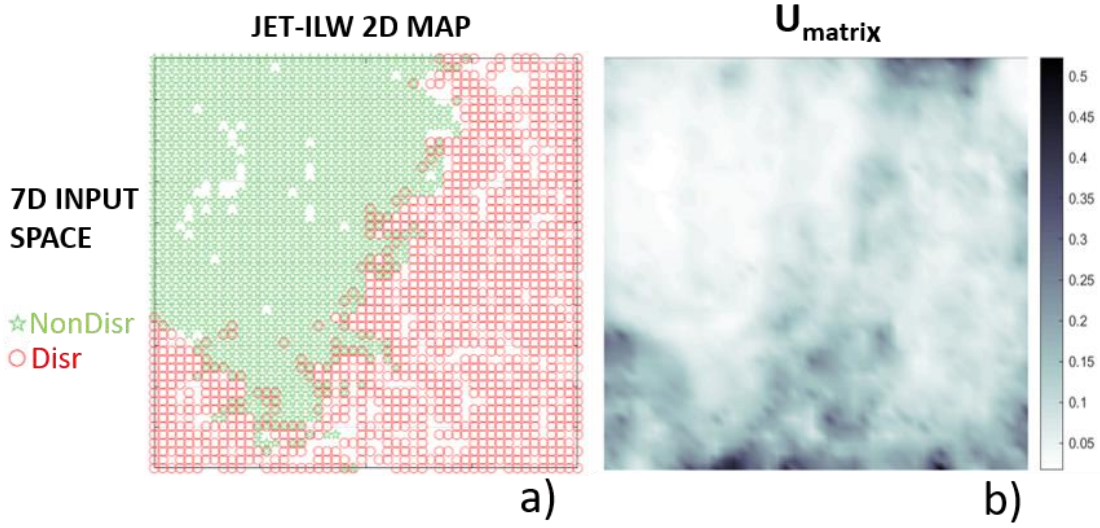

Figure 6 – (a) Mode representation (see paragraph *Generative Topographic Mapping* in section II) of the posterior probability distribution over the latent space for the GTM in Figure 3: green points refers to the regularly terminated discharges, whereas red points refers to the unstable phases of the disrupted discharges; (b) corresponding U-Matrix.

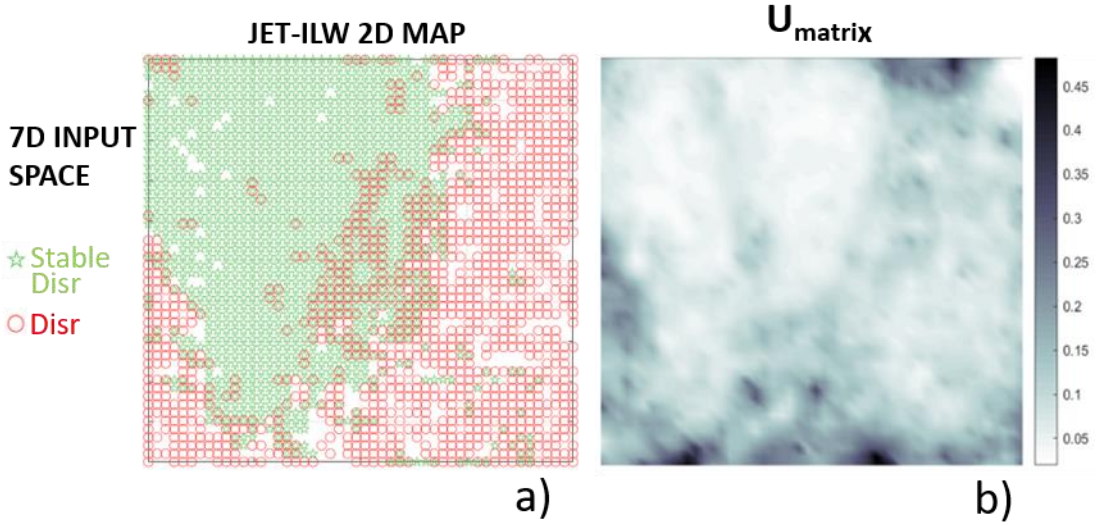

Figure 7 – (a) Mode representation of the GTM trained using only disrupted discharges: green clusters refer to the stable phases and red clusters refer to the unstable phases of the disrupted discharges; (b) corresponding U-Matrix.

V. GTM for Disruption Avoidance

In this section, the potentialities of GTM mapping for the detection of a disruptive behavior early enough to enable avoidance actions are presented.

Tracking of the discharges on the GTM map

In addition to using the GTM map for operational boundaries studies, it can also be used for disruption prediction. In fact, the potentiality of the available toolbox for the GTM [21] allows us to track the temporal sequence of the samples on the map, depicting the movement of the operating point during a discharge. Following the trajectory on the map, it is possible to eventually recognize the proximity to an operational region where the risk of an imminent disruption is high. Based on such GTM information, an alarm would be triggered as will be described in the next subsection. In this work, furthermore, the alarm provided by the pure GTM tracking (*GTM7*) will be studied also in combination with the detection of MHD modes locking (*GTM7+LM*), analyzing the different contributions with respect to the involved time scales. In Figure 8 a), the trajectory of the JET disrupted discharge #82657 is reported, where a gradually changing color scale is used to show the temporal evolution of the discharge, from the lighter in the beginning of the discharge to the darker point that corresponds to the disruption time. As it can be noted, the disruptive discharge starts in a non-disruptive cluster, firstly evolving in the stable (green) region, enters the unstable (red) region, ending in a disruptive cluster, which corresponds to the disruption time. Note that, because of the profile-based indicators defined for the analysis, all the discharges have been projected on the map for the entire time interval of the flat-top phase (up to the disruption time t_D in case of disruptions) where the plasma is in the X-point configuration. In the following, the initial time has been named T_0.

The evolution of the discharge on the map can be better visualized referring to the class membership function shown in Figure 8 b). The class membership function is defined with respect to the composition of the clusters where the operative point is evolving. In particular, it represents the percentage of samples of the considered class in the cluster to which the sample is associated, with respect to the total number of samples in the cluster itself. The experiment referred in Figure 8, performed for fueling and impurity seeding studies with Neon, is characterized by an initial significant Tungsten event at ~13.5s, which perturbs temporarily the seven parameters provided as input to the GTM (Figure 8 c)). After that, the plasma recovers completely up to a second Tungsten event, which again perturbs the parameter space, causing this time a more significant

transition in the class membership functions up to the large influx of W, together with other impurities, in correspondence to the reference warning time T_i, that is slightly before 18s. This latter represents "the point of no return", after which the plasma is definitely destabilized, and we can observe the typical signatures of an impurity accumulation process, with the sawtooth activity stopping and the accumulation of several impurities building up in the plasma core. Because of the strong radiation from the core, as clearly showed by the radiation peaking, the temperature and the current profiles become hollow, as reflected by the temperature peaking factor (Te_{pf}) and the internal inductance (L_i) time evolutions. This condition, where the plasma confinement is already deteriorated, evolves until the locking of an MHD instability, a (m=2, n=1) mode, which eventually triggers the disruption.

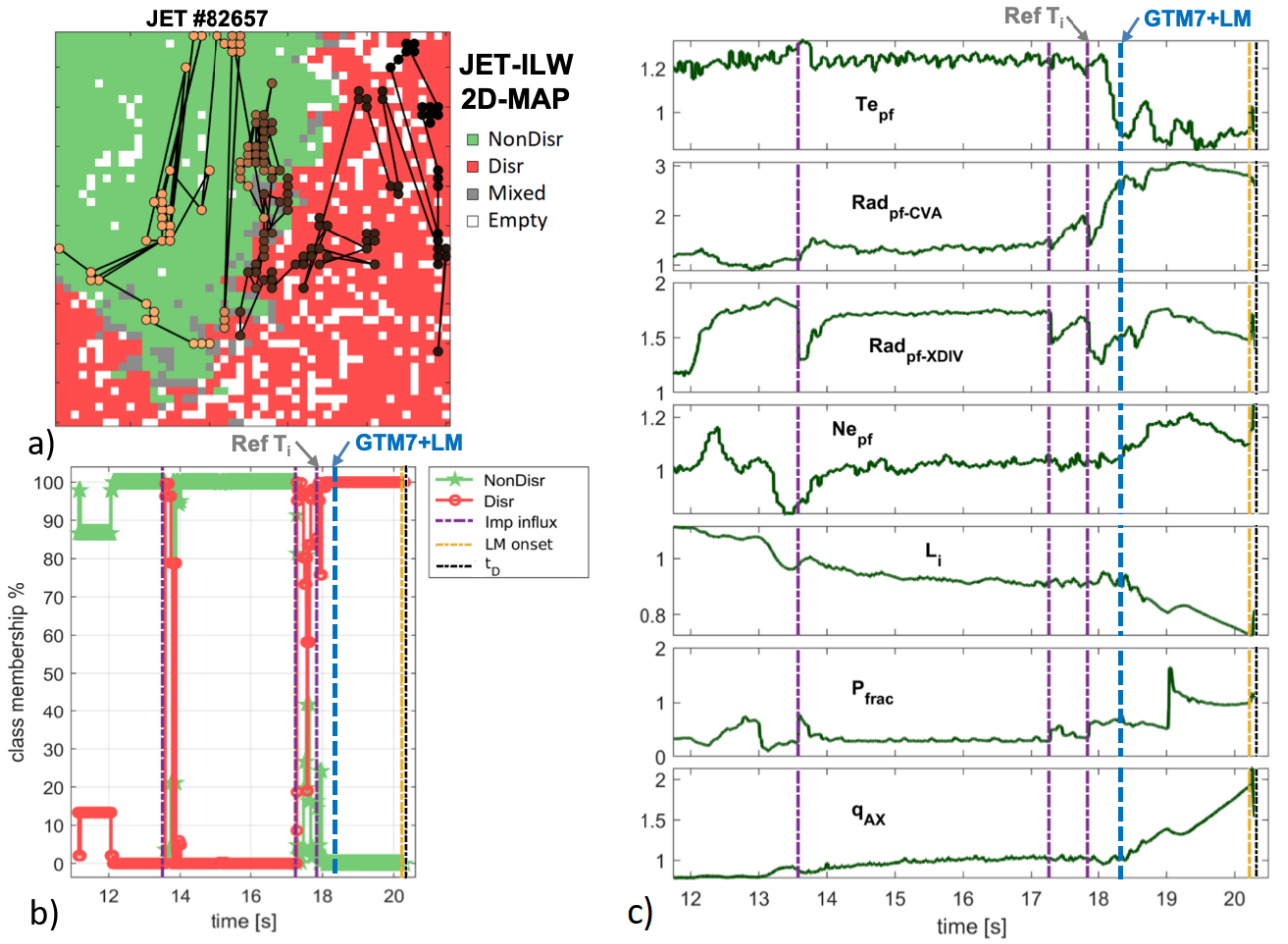

Figure 8 – a) Projection of the disrupted discharge # 82657 on the GTM of Figure 3. The lighter points correspond to the beginning of the discharge, whereas the darker one corresponds to the end, at the disruption time t_D; b) Class member functions of non-disrupted (green) and disrupted (red) classes; c) Time evolution of the 7 plasma parameters used to build the GTM. The vertical dashed lines in b) and c) correspond to specific times of interest characterizing the evolution of the discharge: the time of influx of W and other impurities, *Imp. Influx* (purple), the Reference Warning Time, *Ref* T_i, i.e. the time indicative of the start of the chain of events leading to disruption that has been manually identified (overlapping with the third impurity influx, indicated by a grey arrow); the time where the prediction system would trigger an alarm, *GTM7+LM* (blue); the time of the locked mode onset, *LM onset* (yellow); the disruption time, t_D (black).

As can be noted, the disruption precursors caught by the peaking factors appear well in advance with respect to the onset of the locked mode, corroborating the chance of using these features for a much earlier detection of a disruptive behavior, which is an essential requirement for any action of avoidance.

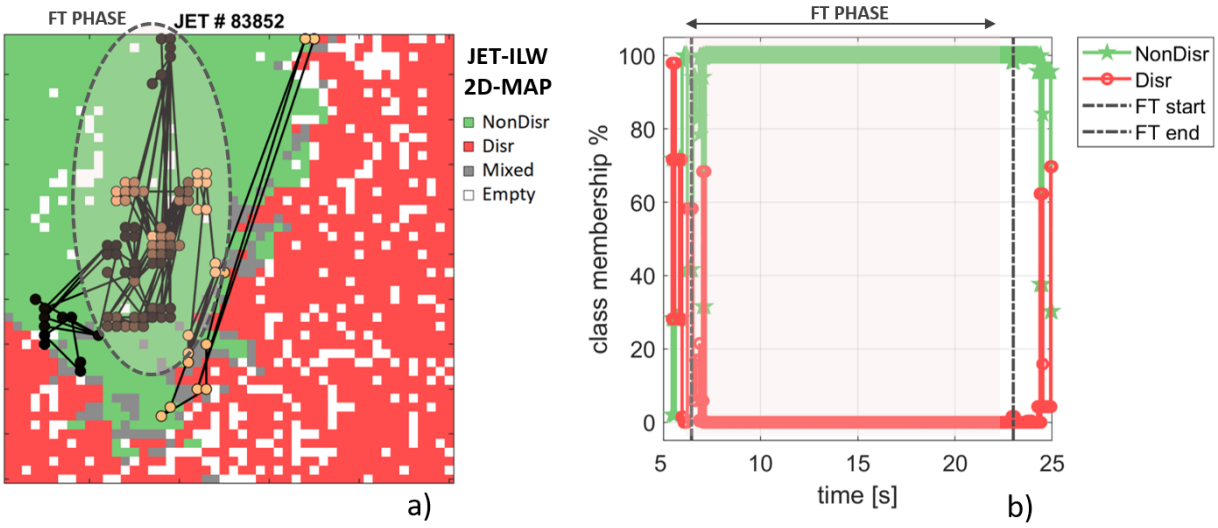

Figure 9 – a) Projection of the non-disrupted discharge # 83852 on the GTM of Figure 3. The color of the circle depicting the movement of the operating point becomes darker and darker as the discharge is approaching to the final phase; b) Class member functions of non-disrupted (green) and disrupted (red) classes.

In Figure 9, the trajectory of the regularly terminated discharge #83852 is reported. As in the majority of the regularly terminated discharges, during the flat-top (FT) phase of the plasma current the non-disrupted discharge trajectory typically evolves within the green "stable" region. It can be noted that in the final ramp-up and in the early ramp-down phases the class-membership functions are slightly perturbed because of the rapidly varying parameters. In this case, the map, beyond the aforementioned "spikey" transitions, is clearly recognizing the non-disrupted nature of the plasma. Any significant change or transition in the plasma state is reflected to some extent in the evolution of the operating point, shifting sometimes the trajectory towards the boundary separating the non-disruptive from the disruptive region on the map. During the time evolution of regular terminations as well as during the stable phases of disruptive discharges, such transitions are mostly localized within the boundary or in a quite narrow layer across the boundary separating the stable from the unstable phases.

Trigger functions and alarm handling

As previously described, following the trajectory on the map throughout the different regions, it is possible to recognize plasma configurations with different risk of disruption. In order to trigger an alarm, a simple criterion has been optimized, which links the disruption risk of the map clusters to the percentage of disrupted samples into the clusters. Moreover, in order to limit possible tardy or missed alarms due to disruptive processes characterized by fast time scales, or false alarms due to transients, the multiple conditions alarm scheme reported in Figure 10 has been implemented to handle the activation of an alarm. In particular, the condition derived from the GTM model is that an alarm is triggered when the trajectory stays in a disruptive or a mixed cluster containing a percentage of disruptive samples $DS>N_1\%$ for at least d consecutive samples. The *delay time* d has been assumed to vary with the time evolution of the discharge with the following exponential law:

$$d = d_1 + \alpha \cdot e^{-\lambda(t-T_0)}$$

where T_0 is the time when the plasma assumes the X-point configuration. Conversely, if the operating point lays in a mixed cluster with a percentage of disruptive samples $N_2\%<DS<N_1\%$, also the Locked Mode amplitude signal, normalized with respect to the plasma current, is considered to trigger the alarm, and a fixed value of the delay time d_2 is assumed. In order to avoid overfitting and maximize the GTM model capability to generalize, the alarm criteria parameters $N_1\%$, d_1, α, λ, and $N_2\%$, d_2, as well as the threshold *Thr* [mT/MA] on the Locked Mode amplitude, have been chosen by optimizing the total prediction error of the GTM on a cross validation set composed by all the training discharges. Note that, this set is completely independent from

the set used to train the GTM model as far as the stable phase of disruptive discharges is concerned, whereas, regarding their unstable phase and the whole phase of non-disruptive discharges, the training set, because of the data reduction, contains only approximately 10% of the total samples. The optimal values correspond to $N_1\%$ =98%, $d_1 = 110$, $\alpha = 300$ and λ=5, $N_2\%$ =50%, $d_2 = 2$, and *Thr*=0.43 mT/MA.
The alarm time of the GTM7+LM disruption predictor is determined by the output of the multiple conditions in the AND/OR logic scheme shown in Figure 10.

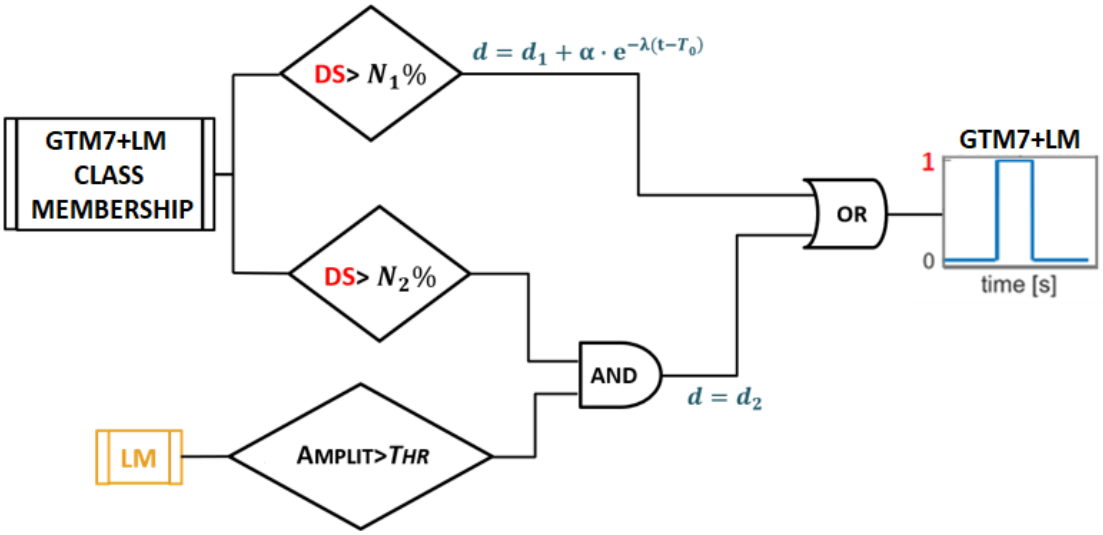

Figure 10 – Multiple condition alarm scheme of the GTM7+LM disruption predictor.

In Figure 11 an example of disruption occurring with a quite fast time scale is shown. The stationarity of the discharge is firstly destabilized by a W event at ~9.8s, followed by an influx of low Z impurities, which through the cooling of the plasma edge leads to the locking of the (*m=2*, *n=1*) mode. In this case, the alarm is triggered by the bottom branch in input to the *OR* logic in the scheme of Figure 10, which is reacting before the top branch. This is a situation occurring mostly when the disruptive process develops on relatively short time scales such that the effect on the 7-D features space, also because of their processing and the delay time *d*, propagates later than the detection of the Locked Mode. Nevertheless, it is worth to note that also in this case there is a well-defined transition in the class-membership functions before the mode locking, with the operating point approaching the boundary between the two classes on the map shortly after the impurity event.

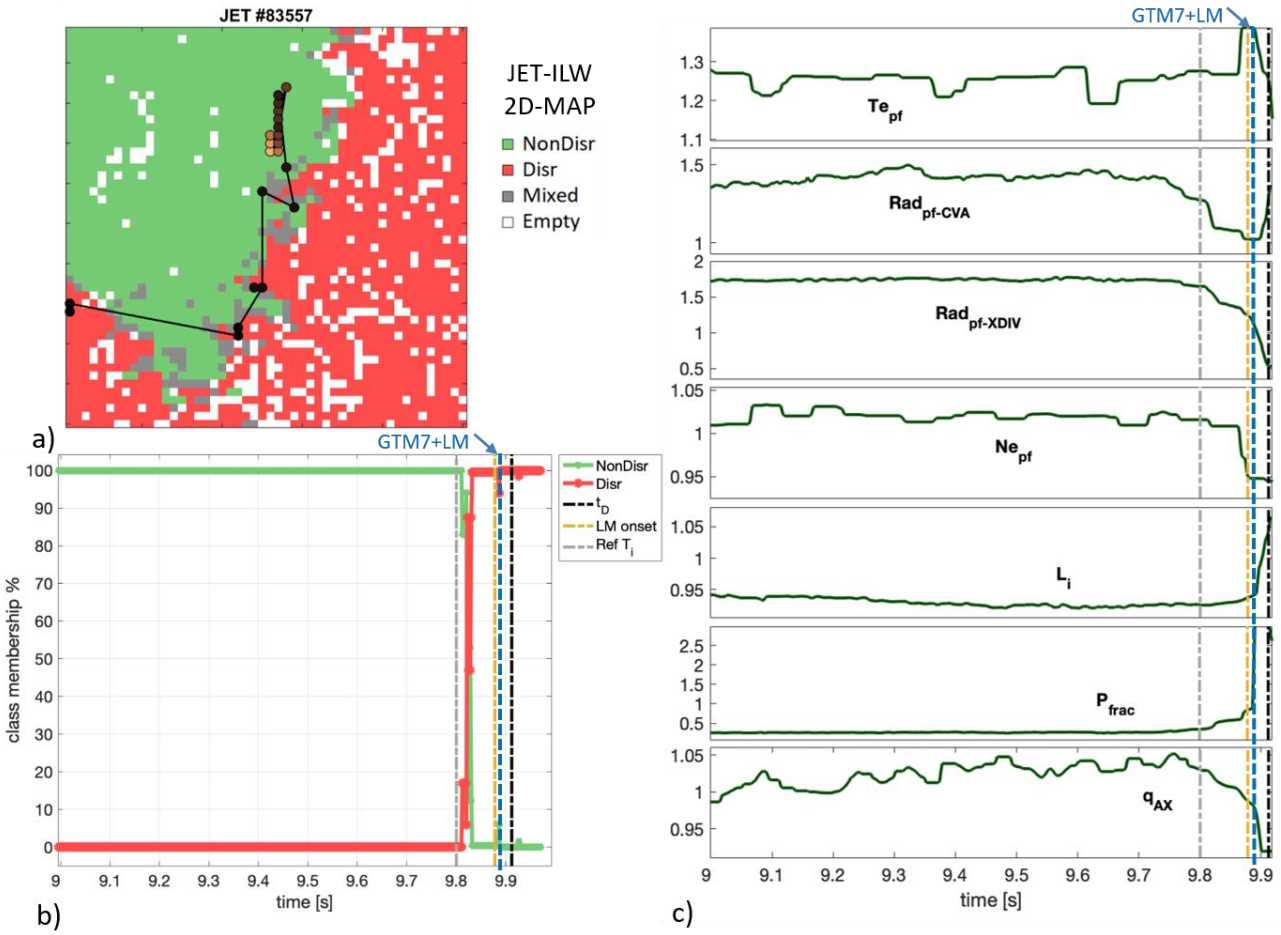

Figure 11 – a) Projection of the final phase of the disruptive discharge # 83557 on the GTM of Figure 3. The color of the circles depicting the evolution in time of the operating point becomes darker and darker as the discharge is approaching to the final phase; b) Class member functions of the non-disrupted (green) and disrupted (red) classes; c) Time evolution of the 7 plasma parameters used to build the GTM. The vertical dashed lines in b) and c) correspond to specific times of interest characterizing the evolution of the discharge: the Reference Warning Time, *Ref* T_i (grey); the time where the prediction system would trigger an alarm, *GTM7+LM* (blue); the time of the Locked Mode onset, *LM onset* (yellow); the disruption time, t_D (black).

Performance of the GTM as Disruption predictor

According to the literature, the performance of the proposed disruption prediction system is evaluated in terms of successful predictions on disruptions (SPs), missed alarms (MAs), tardy detections (TDs) (a detection is considered tardy if the warning time is less than 10 ms), and false alarms (FAs). A sketch to summarize the aforementioned definitions is reported in Figure 12 a). Since the main aim of the present system is the avoidance of disruptions, premature detections are not included in the present analysis but they will be discussed in the next section. The purpose here is to obtain a distribution of the actual warning times as close as possible to the Reference Warning Times T_i, evaluated with respect to the start of the chain of events leading to the disruptions and identified manually.

The performances on the discharges both in the test and in the training set have been evaluated resulting in all correct predictions but one tardive detection (in the training set) and 7 false detections (6%), 3 over the training and 4 over the test datasets. Regarding the "false" detections in regularly terminated discharges, a more in-depth reasoning needs to be done, and this is postponed to the next section about the analysis of the results.

This very good performance is associated to a quite early detection of a disruptive behavior, as shown in Figure 12 b), which reports the cumulative warning time distribution of the proposed GTM7+LM predictor (in blue), evaluated as the difference between disruption time and GTM7+LM alarm time. A second cumulative

warning time distribution, referred to as GTM7 (in cyan), has been reported by neglecting the contribution of the LM alarm. The cumulative warning time distribution reports the fraction of the shots that has a warning time larger than a selected value. The same Figure 12 b) reports the cumulative reference warning time evaluated on the entire dataset (see also Figure 13), that is the difference between disruption time and reference warning time T_i (in grey), and the cumulative Locked Mode time, evaluated as the difference between disruption time and Locked mode onset time (in orange). For all the disruptions, the GTM7+LM alarm warning time is well in advance with respect to the time needed by the DMV to intervene with only one exception (a tardive detection), with more than 55% of the discharges predicted more than 1 second before the disruption time. Furthermore, the GTM7+LM predictor is almost always able to recognize a disruptive behavior well in advance with respect to the Locked Mode predictor.

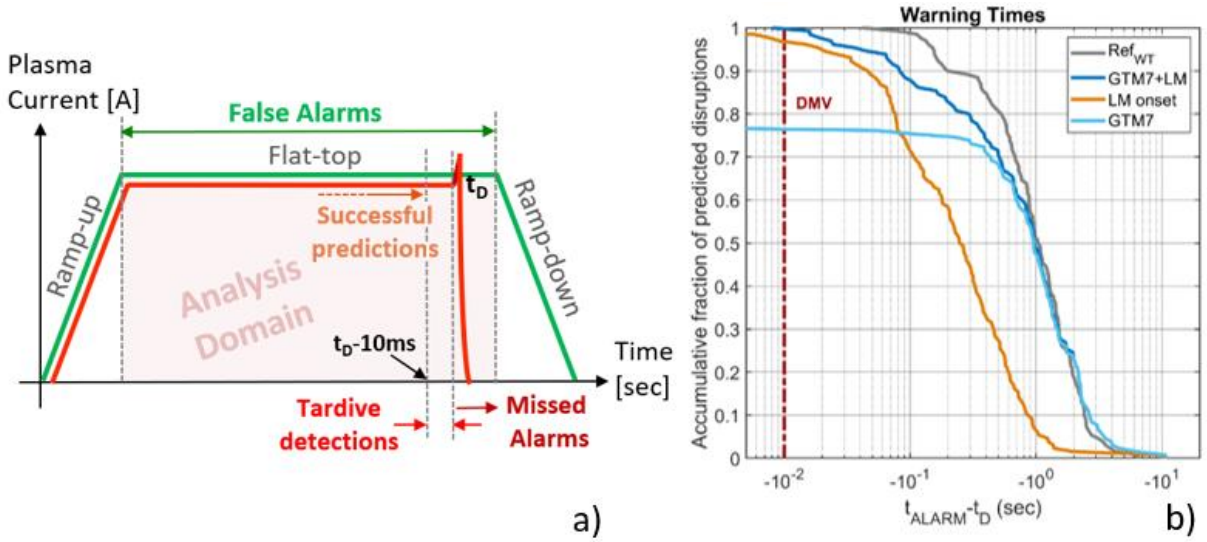

Figure 12 – a) Sketch representing how false alarms, missed alarms, tardive detections and successful predictions are defined; b) Cumulative warning time distributions for all the disrupted discharges in the training and test set of JET (the vertical bar in dark red, *DMV*, allows to identify tardy detections).

By analyzing the cumulative warning times distributions, we can clearly distinguish the two contributions to the global prediction. On longer time scales, the main contribution to the alarm activation is due essentially to the top branch of the alarm scheme in Figure 10: GTM7 and GTM7+LM are mostly overlapping and at around 200ms (that is exactly of the order of the optimized time window during which the alarm condition must persist before the alarm is actually triggered, referred to as assertion time) the GTM7 cumulative distribution is flattening. On the shorter time scales, the predictions are mainly due to the contribution of the locked mode, therefore it's the bottom branch that activates the alarm in those cases. As a further remark about the exploitation of the locked mode information, this scheme is aiming to detect a disruptive behavior with the largest possible warning time since the purpose of the analysis is to avoid disruptions. In many cases, after an initial mode locking, we can have several relaxations and the plasma can survive for hundreds of ms with a small-moderate magnetic island before disrupting. Though as a last "measure of defense", it has been demonstrated on other Tokamaks [30] that it is possible, under certain conditions, to avoid a potential disruption even after the onset of a locked mode by localized ECCD deposition around the q=2 surface, for instance. In some cases, at JET, the discharge was partially recovered from a locked mode, even though at lower performance, by applying ICRH heating, whose optimization is extremely important for core impurity control [31]. Such an early detection of the locking phase would allow at least the possibility of a plasma fast shut-down.

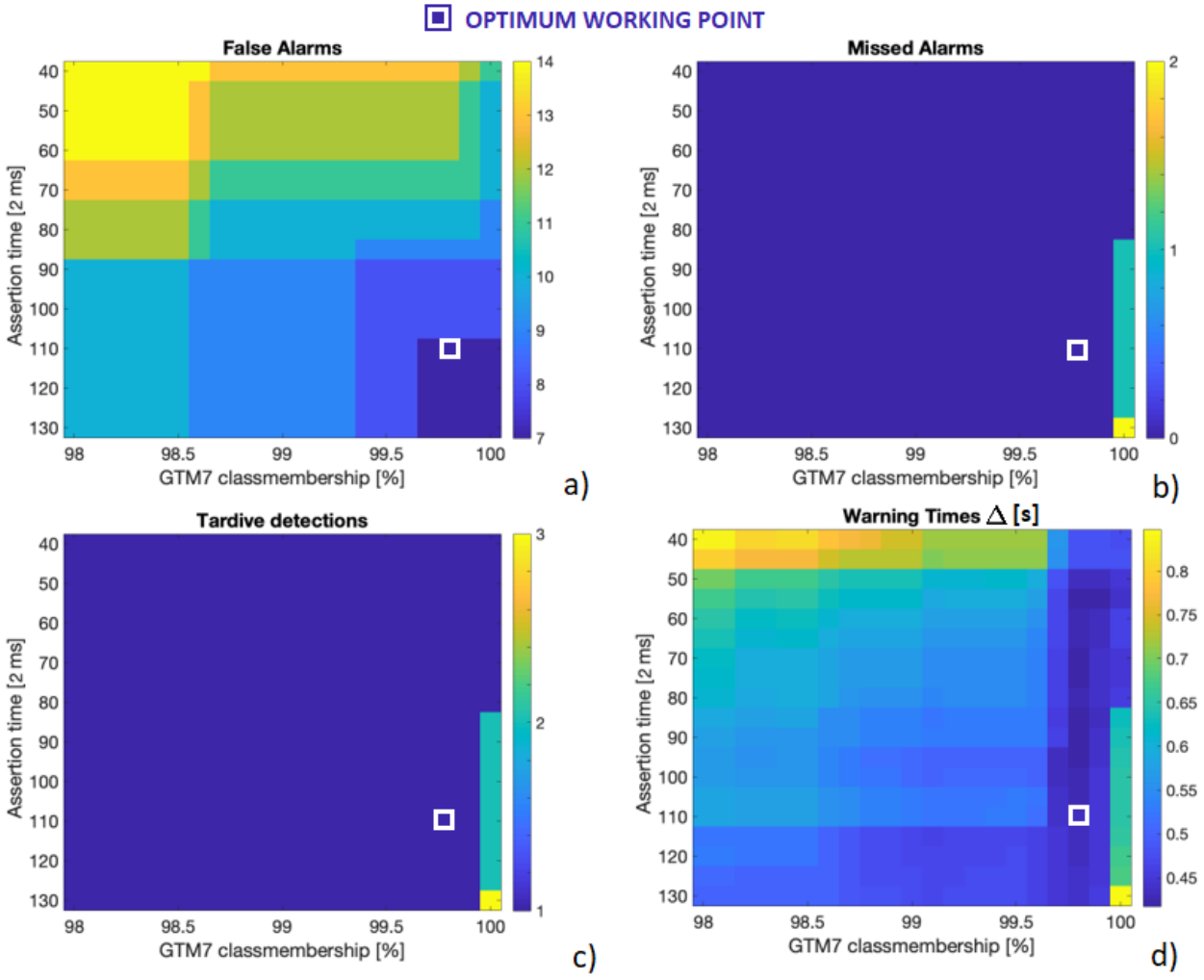

Figure 13 – Distribution of a) false alarms, b) missed alarms, c) tardive detections and d) warning times Δ as a function of the assertion time and the probability (expressed in terms of class-membership percentage) of disruption. The assertion time is defined as the time window during which the alarm condition must persist on a permanent basis before the alarm is actually triggered. Note that the time unit [2ms] has been assumed coincident to the cycle time of the JET real-time network (ATM). The warning times statistical dispersion is defined as the mean absolute deviation between Ref T_i and the alarm triggered according to the scheme described in Figure 10.

The cumulative distributions of the warning times previously described correspond to the optimal working point reported on Figures 13 a), 13 b) ,13 c) and 13 d). These figures represent, respectively, the distribution of false alarms, missed alarms, tardive detections and warning times Δ as a function of the assertion time and the probability (expressed in terms of class-membership percentage) of disruption. Given the target of the analysis (disruption avoidance), the high separability between the disruptive and the non-disruptive class and the need to filter out transient phenomena, in order to improve the performance, we have to move on the diagonal from the left-upper corner to the right-bottom corner. The distributions of the considered performance indicators are not only consistent one another, but, moreover, their variation is well-defined and smooth. This is indicative of the robustness of the information that can be extracted with this analysis in the considered parameters space.

Another important outcome of the analysis is in relation to the assertion time, that compared to previous disruption predictors discussed in literature (which however were not designed for avoidance purposes), has to be much larger. This is consistent with the longer time scales (such as the transport time scales associated to the accumulation of impurities) required for the physics mechanisms described in this paper to take place.

Analysis of the results

As presented in the previous section, the proposed system performs very well in terms of successful prediction with, given the large warning time, a limited number of false alarms in the considered JET-ILW campaigns (7 in the optimum working point, that corresponds to 6% of the entire dataset). The definition of false alarms has been kept as a legacy of previous disruption prediction studies, but in this context its interpretation is clearly different and will be discussed more in detail in the following.

Another figure of merit, which is often taken into account to evaluate the performance of a prediction system, is the rate of premature detections, even if this assessment has always been done considering a fixed threshold determined on statistical basis (a typical value in the case of JET was 2.5 seconds). On the contrary, being the range of the involved time scales quite large, a fixed threshold does not allow to define a good indicator for premature detection. The most reasonable way to assess the consistency of the analysis from this point of view is to evaluate how close the triggered alarms are with respect to what has been identified as the reference warning time T_i. In most of the cases, the time of the alarm and the corresponding reference time are quite close, as can be seen statistically by looking at the differences between the cumulative warning times represented in Figure 12. In some cases, there exist unstable phases followed by relatively partial recovers of the plasma, with the alarm triggered on the first "event"; there are also cases where the transition in the considered 7-D feature space is delayed because of different reasons, as discussed throughout section V.

The non-disruptive discharge reported in Figure 14 shows one of the 7 "false alarms". After the switch-off of the ICRH heating (slightly before 17s) the influx of several impurities degrades plasma energy confinement causing the cooling of the edge by radiative collapse. In this phase the trajectory of the operating point cross the disruptive boundary on the map (after 17.5s) after evolving across the boundary and penetrating for short phases well inside the disruptive region on the bottom-left corner of the map. The discharge exhibits even a locked mode persisting almost 1s, promptly detected by the JET control system that react with a soft stop and is then able to safely land the plasma current. In this case the alarm has been triggered by the bottom branch of the scheme (depending on the locked mode), but it is worth noticing that a potential disruptive behavior is clearly detected also by the projection on the map. The resulting pattern is clearly corresponding to an edge radiative collapse (see Figure 15, where within the unstable phase of disruptive discharges, the contribution of edge (EdgeRC) and core (CoreRC) radiative collapses has been distinguished [3]), as can be easily extrapolated by analyzing the component planes of the input features space (see Figure 4 and 5). Even if the plasma current was safely landed, the detection is without any doubt consistent with a potential disruptive condition from which the plasma has to be recovered and that requires either an emergency response or a reaction by the control system. The response of this latter, moreover, has modified the dynamics of the discharge and post-facto it is difficult to state if the discharge would have survived without any intervention.

In the context of disruption avoidance and plasma termination, the possibility of recognizing different patterns corresponding to specific sequence of events leading to disruption would allow the optimization of more targeted responses by the control systems. Even if a discussion about the implementation of dedicated control strategies for this aim is beyond the scope of the work, it is worth to spend some words about the implications of having the possibility to discriminate among different disruption causes.

As previously mentioned with reference to Figure 15, the considered physics-based indicators allow the identification of two clearly distinct patterns for Edge and Core radiative collapses. Such a representation has been obtained distinguishing with a different label the unstable phases of disruptive discharges characterized either by one or by the other radiative collapse. The identification in a complex high-dimensional space of patterns corresponding to different physics mechanisms was one of the missing pieces in the view of an integration of data-driven models into disruption avoidance strategies. This analysis clearly shows that, given proper physics-based indicators, a proper representation of the structure of the data and an accurate characterization of the disruptive unstable phase, data-driven predictions can be "physics-driven" as well. This extremely relevant aspect was already introduced in [3] and will be the object of more detailed analyses for the practical integration of this approach in PETRA.

Summarizing the outcome of the analysis, out of 7 “false” detections, in 2 cases (in the test set) the alarm is triggered by the bottom branch because of a well-developed locked mode and, in 5 cases (all of which after either the step-down or the switch-off of the heating power), the alarm would be however triggered because of the appearance of a locked mode. In 6 cases out of 7, the alarm would be triggered by the top branch because of a radiative collapse strongly degrading plasma confinement. All the detections are consistent with a potential disruptive condition, characterized by the presence of a large N=1 mode slowly rotating or locked, with a plasma poorly performing which requires an intervention.

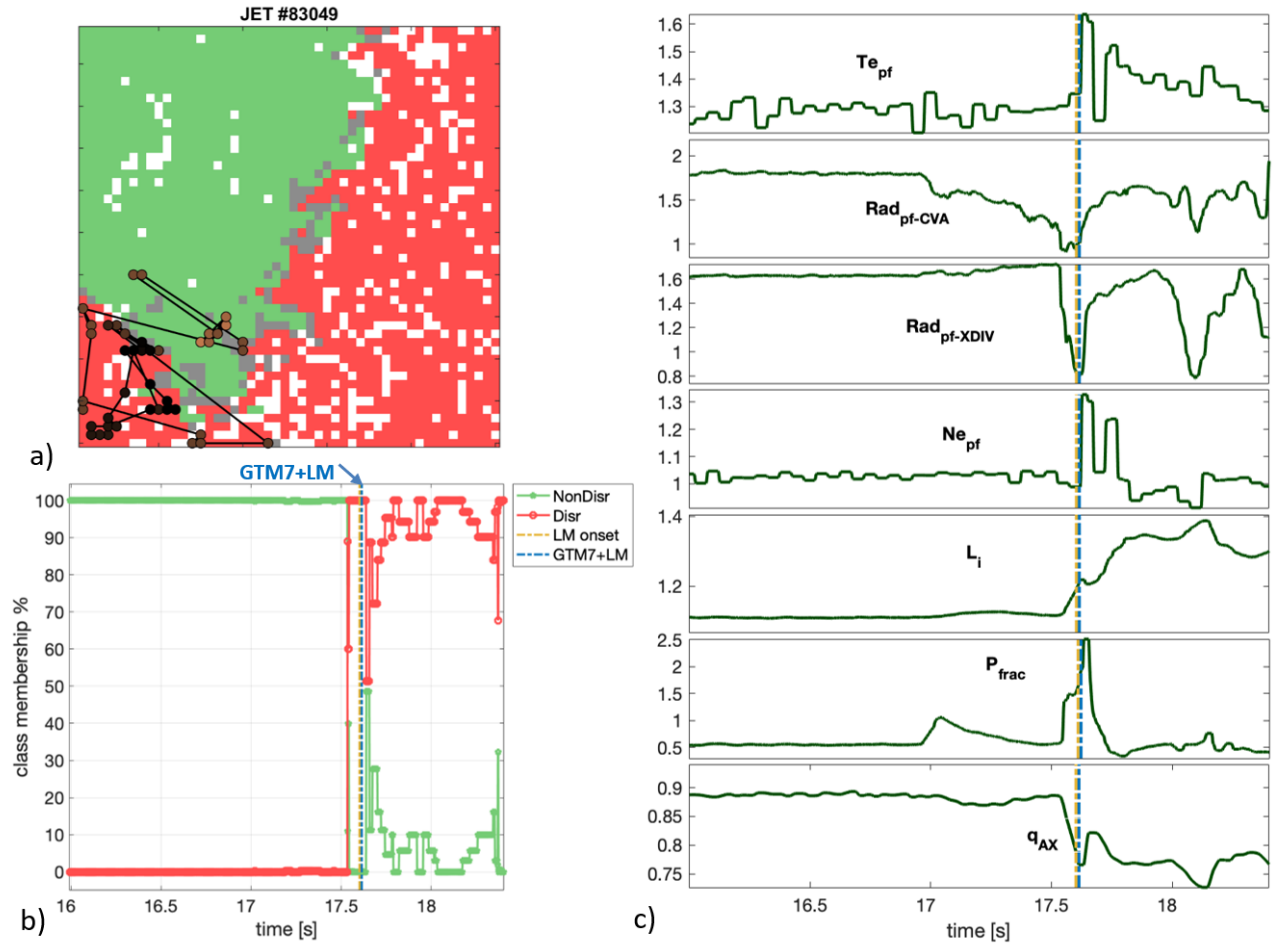

Figure 14 – a) Projection of the non-disrupted discharge # 83049 on the GTM of Figure 3. The color of the circles depicting the evolution in time of the operating point becomes darker and darker as the discharge is approaching to the final phase; b) Class member functions of non-disrupted (green) and disrupted (red) classes; c) Time evolution of the 7 plasma parameters used to build the GTM. The vertical dashed lines in b) and c) correspond to specific times of interest characterizing the evolution of the discharge, in this case the time of the locked mode onset, *LM onset* (yellow) and the time *GTM7+LM* (blue) corresponding to the alarm triggered according to the scheme in Figure 10.

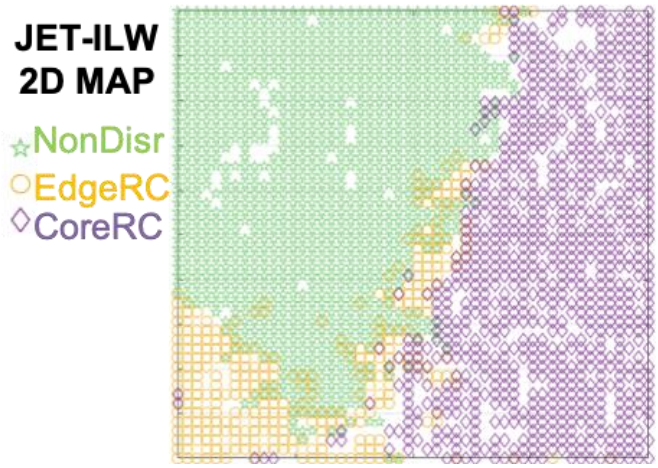

Figure 15 – a) GTM 2D map reporting the projection of the modes of the posterior probability distribution. Within the unstable phase of disruptive discharges, the contribution of edge (*EdgeRC*) and core (*CoreRC*) radiative collapses has been distinguished. Consistently to what has been found in [3], the two macro-classes of disruption are described by quite different patterns on the map and this is a fundamental prerequisite for planning proper avoidance strategies.

VI. Conclusions

It is worth emphasizing that, compared to other disruption prediction approaches belonging to machine learning techniques, such as neural networks, the GTM provides significant additional value. Whereas neural networks are "black boxes", which provide a prediction but are very difficult to interpret, on the contrary, as previously shown, the map allows to follow the trajectory of the plasma in the parameter space of interest and to study its behavior leading to a disruption. Thus, the developed map has the potentiality to provide much more than a simple prediction in the understanding of the operational space and the causes of the disruptions. Provided a suitable parameter space, the evolution of the operating point on the map can be characterized according to recurrent patterns describing different mechanisms that cause the destabilization of the discharge. The analysis of those patterns, especially if the manifold is not excessively complex, allows studying the combination of parameters associated to different regions of the operational space and how their combination correlates with different paths leading to disruptions. This aspect is extremely valuable from the point of view of the interpretative analysis and is defining a new paradigm in many fields. In relations to these points, the effectiveness of the approach and the tool developed for the analysis should be emphasized. One of the most challenging things in a field as complex as the control of fusion plasmas is the capability to extract useful and robust information from physics quantities, which either represent or are directly connected to "observables" that can be controlled to act on the plasma state. Profile based indicators, even being only one of the ingredients that play a role on plasma performance and stability, go without doubt in this direction. Concerning the tool developed for the analysis, the GTM framework allows to efficiently implement a non-trivial machine learning workflow, incorporating pattern recognition and classification algorithms, and, computationally, is well capable of satisfying real-time requirements.

All these elements, together with the possibility to have warning times compatible with disruption avoidance, make this approach unique in the plethora of those proposed so far to identify a disruptive behavior. The considered indicators are based on dimensionless physics quantities, in principle routinely available on most of the machines.

Nevertheless, it is worth mentioning a not negligible drawback linked to some of the considered features, in relation to the signals required to compute them presently available in real-time. For instance, ECE measurements were totally replaced by HRTS measurements because of the not negligible effect on the electron temperature peaking factor calculation. In particular, cut-offs propagating on a significant number of channels resulted in artifacts and preferential unwanted patterns distorting the tracking of the time evolution on the map, especially in correspondence of the transitions between stable and unstable phase. Nevertheless, the present work has to be considered a proof of concept of the potentiality of the proposed machine learning

workflow for disruption avoidance. The assessment of the reliability of measurements during real-time operations is a serious concern which deserves careful considerations, but it is beyond the scope of this work. Presently, strategies to deal with potential problems affecting measurements and "backup solutions" are under investigation and will be the object of future studies.
Another well-known concern, more related to the exploitation of machine learning techniques for disruption prediction, is the capability to extrapolate and generalize to different scenarios and different machines, which is common to any data-driven approach applied outside the training domain. Nevertheless, the choice of dimensionless parameters, reflecting physics underlying mechanisms, as well as very preliminary analyses on ASDEX Upgrade, have shown the possibility to identify similar patterns. There are other important parameters whose integration in the analysis might have a beneficial effect, such as MHD rotating modes. Their activity, as well-known, plays a key role in the plasma stability and, in case of tearing modes, develops on time scales similar to those of the considered plasma profiles [32]. However, supplementing the information associated to rotating modes without affecting the robustness of the developed system is all but an easy task and is already undergoing significant efforts.
Concluding, the analysis presented in this paper demonstrates how machine learning tools can efficiently deal with the information extracted from structured data (1-D profiles), and how this "knowledge" can be robustly exploited for disruption prediction and avoidance. This is a very promising starting point, dispelling the myth that any machine learning tool is necessarily a black box, which cannot provide any useful insight on main physics mechanisms leading to disruptions. Machine learning is an extraordinary resource that, if properly used, can help to take important steps forward in fusion as it is happening in many other fields of science.

Acknowledgments

This work has been carried out within the framework of the EUROfusion Consortium and received funding from the EURATOM research and training programme 2014–2018 and 2019-2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

References

[1] N.W. Eidietis , W. Choi, S.H. Hahn, D.A. Humphreys, B.S. Sammuli, M.L. Walker, (2018), "Implementing a finite-state off-normal and fault response system for disruption avoidance in tokamaks", Nuclear Fusion, 58, 056023.

[2] M. Lehnen, et al., (2015), "Disruptions in ITER and strategies for their control and mitigation, Journal of Nuclear Materials, 463, 39–48.

[3] A. Pau, A. Fanni, B. Cannas, S. Carcangiu, G. Pisano, G. Sias, P. Sparapani, M. Baruzzo, A. Murari, F. Rimini, M. Tsalas, P.C. de Vries, (2017), "A first analysis of JET plasma profile-based indicators for disruption prediction and avoidance", IEEE Transactions on Plasma Science, DOI:10.1109/TPS.2018.2841394.

[4] P. C. de Vries, G. Pautasso, D. Humphreys, M. Lehnen, S. Maruyama, J. A. Snipes, A. Vergara, and L. Zabeo (2016), "Requirements for Triggering the ITER Disruption Mitigation System", Fusion Science and Technology, 69, 471-484.

[5] G. Pautasso, et al. (2018), "The ITER disruption mitigation trigger: developing its preliminary design", Nuclear Fusion, 58, 036011.

[6] B. Cannas, A. Fanni, P. Sonato, K. Zedda, (2007), "A prediction tool for real-time application in the disruption protection system at JET", Nuclear Fusion, 47, 11, 1559-1569.

[7] G.A. Rattá, J. Vega, A. Murari, G. Vagliasindi, M.F. Johnson, P.C. de Vries, (2010), "An advanced disruption predictor for JET tested in a simulated real-time environment", Nuclear Fusion, 50, 025005.

[8] B. Cannas, A. Fanni, G. Pautasso, G. Sias, (2011) "Disruption prediction with adaptive neural networks for ASDEX Upgrade", Fusion Engineering and Design, 86, 6-8, 1039-1044.

[9] B. Cannas, A. Fanni, G. Pautasso, G. Sias (2010), "An adaptive real-time disruption predictor for ASDEX upgrade", Nuclear Fusion, 50, 7, 075004.

[10] W. Zheng et al (2018), "Hybrid neural network for density limit disruption prediction and avoidance on J-TEXT tokamak", Nuclear Fusion, 58, 056016.

[11] C. Rea, R. S. Granetz (2018), "Exploratory Machine Learning Studies for Disruption Prediction Using Large Databases on DIII-D", Fusion Science and Technology, ISSN: 1536-1055 (Print) 1943-7641 (Online).

[12] Dormido-Canto, J. Vega, J.M. Ramírez, A. Murari, R. Moreno, J.M.López, A. Pereira, (2013), "Development of an efficient real-time disruption predictor from scratch on JET and implications for ITER", Nuclear Fusion, 53, 113001.

[13] B. Cannas, P. De Vries, A. Fanni, A. Murari, A. Pau, G. Sias, (2015), "Automatic disruption classification in JET with the ITER-like wall", Plasma Physics and Controlled Fusion, 57, 12, 125003.

[14] B. Cannas, A. Fanni, A. Murari, A. Pau, G. Sias, (2013) "Automatic disruption classification based on manifold learning for real-time applications on JET", Nuclear Fusion, 53, 9, 093023.

[15] R. Aledda, B. Cannas, A. Fanni, A. Pau, G. Sias, (2015) "Improvements in disruption prediction at ASDEX Upgrade", Fusion Engineering and Design, 96-97, 1 698-702.

[16] V.N. Vapnik (1995) The Nature of Statistical Learning Theory. New York: Springer-Verlag.

[17] T. Kohonen (1998), "The self-organizing map", Neurocomputing, 21, 1–3, 1-6.

[18] C. M. Bishop, M. Svensen (1998), "The Generative Topographic Mapping", Neural Computation, 10, 1, 215-234.

[19] S. Ben-David, S. Shalev-Shwartz. (2014) Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press.

[20] Murari A. et al (2013) "Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions", Nuclear Fusion, 53, 033006.

[21] A. Pau, "Techniques for prediction of disruptions on TOKAMAKS", Ph.D. dissertation, 2014, Centro Interdipartimentale "Centro Ricerche Fusione", Padova. [Online] available: http://paduaresearch.cab.unipd.it/6664

[22] S.T. Roweis, L. K. Saul, (2000), "Nonlinear dimensionality reduction by locally linear embedding"., Science 290 (5500), 2323–2326.

[23] J.A. Lee, M. Verleysen, (2007), "Nonlinear dimensionality reduction", Springer.

[24] C. M. Bishop, (2006), "Mixture models and the EM algorithm", Available online at http://www.cs.ubbcluj.ro/ csatol/gep-tan/Bishop-CUED-2006.pdf.

[25] A. Pau, B. Cannas, A. Fanni, G. Sias, M. Baruzzo, A. Murari, G. Pautasso, M. Tsalas, (2017) “A tool to support the construction of reliable disruption databases”, Fusion Engineering and Design, 125, 139-153.

[26] L. Barrera et al. (2010), “Inboard and outboard electron temperature profile measurements in JET using ECE diagnostics”, Plasma Physics and Controlled Fusion, 52 085010.

[27] A. Huber, et al. (2007), “Upgraded bolometer system on JET for improved radiation measurements”, Fusion Engineering and Design, 82, 1327–1334.

[28] S. Carcangiu, A. Fanni, A. Montisci, (2008) “Multiobjective tabu search algorithms for optimal design of electromagnetic devices, IEEE Transactions on Magnetics, 44, 6, 4526820, 970-973.

[29] P. C. de Vries, et al. (2014), “The influence of an ITER-like wall on disruptions at JET”, Physics of Plasmas, 21, 056101.

[30] M. Maraschek et al. (2018) “Path-oriented early reaction to approaching disruptions in ASDEX Upgrade and TCV in view of the future needs for ITER and DEMO”, Plasma Physics and Controlled Fusion 60, 014047.

[31] E. Lerche et al. (2016) “Optimization of ICRH for core impurity control in JET-ILW”, Nuclear Fusion 56, 036022

[32] C. Sozzi, E. Alessi, M. Baruzzo, S. Gerasimov, (2016) “Development of a disruption precursor based on rotating MHD instabilities and its application to JET H-mode plasma scenario”, 58th Annual Meeting of the APS Division of Plasma Physics, Volume 61, Number 182016; San Jose, California.