Progress Towards Interpretable Machine Learning-based Disruption Predictors Across Tokamaks

C. Rea[*,+,1], K.J. Montes[+,1], A. Pau[+,2], R.S. Granetz[1], and O. Sauter[2]

[1] *Massachusetts Institute of Technology (MIT)*
Plasma Science and Fusion Center (PSFC), Cambridge, MA, USA
[2] *Ecole Polytechnique Fédérale de Lausanne (EPFL),*
Swiss Plasma Center (SPC), CH-1015 Lausanne, Switzerland

[+]These authors have equally contributed to the realization of this manuscript.

[*]Email: crea@mit.edu

[*]Mailing Address: *General Atomics, G13-366*
3483 Dunhill Street, San Diego, CA 92121

[*]Phone: +1 858 455 2157

2019-IAEA-FDPVA Special Issue

Number of pages: 27

Number of tables: 2

Number of figures: 9

Abstract

In this paper we lay the groundwork for a robust cross-device comparison of data-driven disruption prediction algorithms on DIII-D and JET tokamaks. In order to consistently carry on a comparative analysis, we define physics-based indicators of disruption precursors based on temperature, density, and radiation profiles that are currently not used in many other machine learning predictors for DIII-D data. These profile-based indicators are shown to well-describe impurity accumulation events in both DIII-D and JET discharges that eventually disrupt. The univariate analysis on the features used as input signals in the data-driven algorithms applied on both tokamaks data statistically highlights the differences in the dominant disruption precursors. JET with its ITER-like wall is more prone to impurity accumulation events, while DIII-D is more subject to edge cooling mechanisms that destabilize dangerous MHD modes. Even though the analyzed datasets are characterized by such intrinsic differences, we show through few examples that the inclusion of physics-based disruption markers in data-driven algorithms is a promising path toward the realization of a uniform framework to predict and interpret disruptive scenarios across different tokamaks. As long as the destabilizing precursors are diagnosed in a device-independent way, the knowledge that data-driven algorithms learn on one device can be re-used to explain a disruptive behavior on another device.

Keywords — Disruptions, Machine Learning, DIII-D, JET.

I. INTRODUCTION

As the fusion community is focusing to develop intrinsically stable scenarios for reliably safe operations, disruptions in tokamak plasmas still pose serious challenges. The sudden loss of plasma energy and its confinement can cause deleterious damage to plasma facing components, as well as imposing mechanical stresses on the device. In view of future devices, like ITER [1] or SPARC [2], for which such generated forces will represent an unbearable obstacle to safe operations, disruption mitigation and avoidance has become an active and pressing area of research [3–5]. The statistical analysis of disruptive instabilities reported in [6] has also enabled many data-driven applications [7–10] to predict disruptions with enough warning time to more efficiently enable avoidance strategies, and some of these solutions [11–13] are also being developed to operate in the Plasma Control System (PCS). Nevertheless, the goal of disruption prediction research is not only to provide predictions of impending disruptions early enough but also to inform the PCS on the offending feature(s), or plasma descriptors, in order to steer the plasma away from the disruptive boundary. For example, if shaping parameters are found to be dangerously contributing to push the plasma toward a disruptive scenario, a continuous monitoring of such drivers through PID controllers can stabilize and optimize plasma behavior. Identifying the actual causes of disruptions and therefore their precursors, is the ideal path toward a successful extrapolation to future devices. To this aim, it is extremely important to focus on uniform physics-based markers when developing data-driven applications, i.e. adopting dimensionless or device-independent plasma descriptors as input features in our models enables more reliable domain adaptation and knowledge extraction across tokamaks.

In this work we aim at reporting some preliminary analysis that include 1D/2D profile information to identify earlier precursors in the disruptive chain of events. Contrary to deep learning strategies for feature extraction from profiles using Convolutional Neural Networks [14, 15], we focus on more classical feature engineering. Deep learning methods may be attractive for they tend to generalize better and have shown highly accurate predictive capabilities, but it comes at a high cost: the accessible interpretability of the model and the features learnt by it. Even though some techniques do exist to try and interpret neural networks, such as sensitivity analysis [16] or axiomatic attribution [17] methods, these are notoriously harder than other classical machine learning

approaches. By imposing (or engineering) the characteristics of the features that the model will learn, we can enhance its interpretability and at the same time connect the model's predictions with the disruptive precursors dynamics. Therefore, following the method outlined in [18], we map the profile diagnostics onto flux surfaces or specific core/edge/divertor regions to reduce the feature dimensionality and obtain peaking factors to use in data-driven disruption prediction algorithms. The peaking factors obtained in this way can be regarded as device-independent indicators and are extremely suitable for cross-device analysis, as they are able to isolate the physics of interest independently from the type of diagnostics or geometry of the device.

The paper is structured as follows: in Section II we discuss the peaking factors and the methodology applied to synthesize them from DIII-D data and diagnostics. Then in Section III we show that the univariate analysis on the most relevant features, such as the temperature profile peaking factor and the Greenwald density fraction, reveals intrinsic differences between DIII-D and JET precursors' dynamics. Section IV discusses the first applications of classical machine learning predictors such as Random Forest (RF) and Generative Topographic Mapping (GTM), trained on DIII-D and tested on JET discharges and vice-versa. In particular, we discuss the predictive output of these data-driven models can be interpreted, also thanks to the information given by the profile peaking factors, thus connecting machine learning predictors with the underlying disruptive mechanisms. Finally, in Section V we draw our conclusions and lay out the future research directions.

II. PEAKING FACTOR ENGINEERING

Many data-driven analyses in fusion research tend to use databases of 0D parameters to make inferences about the plasma state. However, higher dimensional information is often relevant to the problem and ought to be included in some way. In the case of disruption prediction, changes in kinetic and radiative profiles are often connected with the development of destabilizing physics mechanisms, e.g. magnetohydrodynamic (MHD) precursors [19–21]. In order to include the radial profiles of T_e, n_e, and P_{rad} in data-driven analyses, 0D *peaking factor* (PF) metrics have been synthesized on DIII-D following the method outlined in [18]. The PFs for T_e and n_e compare the average value in the plasma core to the average value over the entire profile. Two additional

PFs for P_{rad} are defined so as to decouple contributions from the core and divertor regions of the plasma.

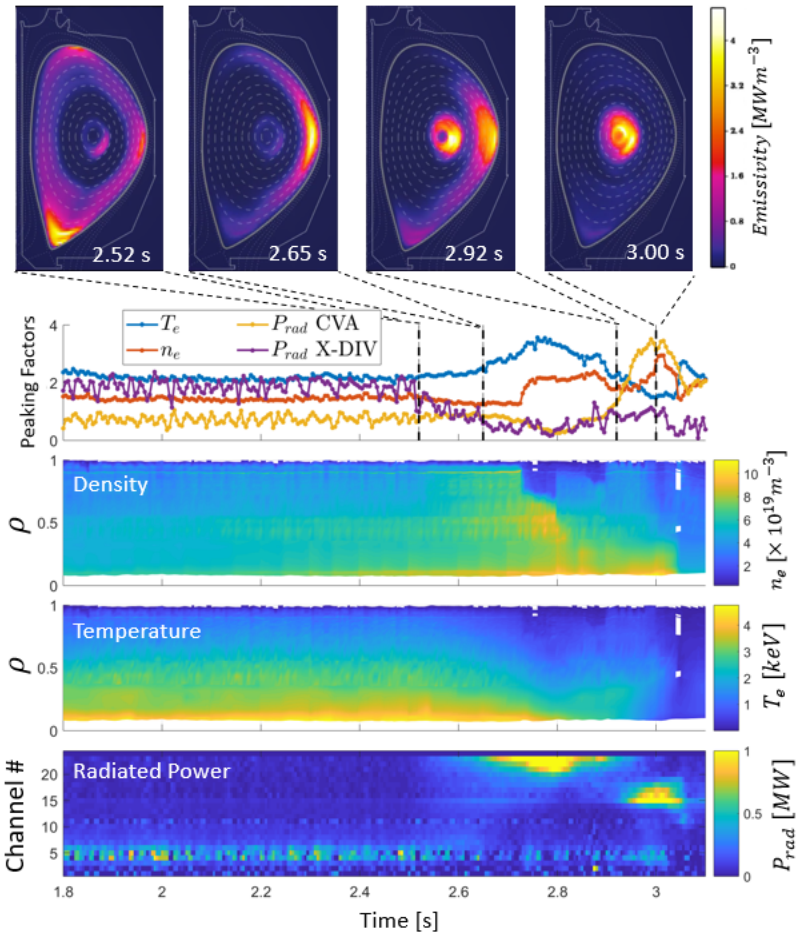

Fig. 1. Peaking factor and profile evolution during DIII-D shot 175697, with tomographic reconstructions of bolometer emissivity at 4 different times of interest; in the bottom three panels, ρ refers to the normalized effective radius given in Eq. 6 and the channel number is that of the lower fan as shown in Fig. 2.

In Figure 1 we show an example of the evolution of all 4 peaking factors for a specific DIII-D discharge. For this particular discharge, an impurity concentration was injected at $\sim 2.5\ s$, causing an increase in P_{rad} at the plasma edge and a corresponding decrease in the P_{rad} divertor peaking factor. As power is radiated away, the edge begins to cool and the temperature profile becomes more peaked, causing an increase in the T_e peaking factor. There is then a transition from H-mode to L-mode at $\sim 2.73\ s$ as the n_e pedestal is lost, causing a change in the baseline of the n_e peaking factor. Finally, starting at $\sim 2.9\ s$ the impurities penetrate into the core and radiate power there,

correlating with an increase in the P_{rad} core peaking factor and a decrease of the T_e peaking, signaling a central cooling.

The calculation of each of these peaking factors relies on a set of diagnostics that can frequently and robustly measure T_e, n_e, and P_{rad} for a large number of shots. Since the types of diagnostic systems available and their attributes vary amongst devices, the peaking factor definitions must be changed accordingly. In the following sections, the methodology used to calculate each of these peaking factors on DIII-D is explained in detail.

II.A. P_{rad} Peaking Factor Calculations

Measurements of the radiated power distribution on DIII-D are acquired using two bolometer arrays consisting of 24 channels each, as shown on the left in Fig. 2. In order to define a 'core vs all' (CVA) and 'divertor vs all' (X-DIV) peaking factor for P_{rad} following a similar method to that used in [18], we restrict our analysis to lower single null plasmas (where the active X-point is on the lower divertor) and use the lower fan as shown on the right in Fig. 2. Nevertheless, this methodology can be extended to upper single null and double null configurations as well.

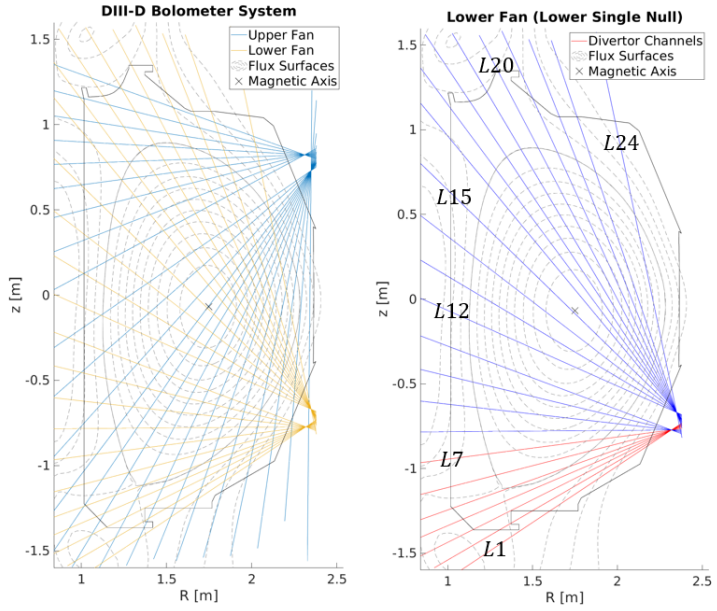

Fig. 2. (Left) Bolometer arrays on DIII-D; (Right) Lower fan channel setup, with divertor channels $L1 - L7$ shown in red.

As described in [22], the power P_j radiated along the viewing chord of channel $j \in [L1, \ldots, L24]$ is calculated by scaling the power radiated onto the detector by a geometric factor specific to that channel. These measurements are then divided into bins based on the region of coverage of the corresponding channel. For the divertor bin D, channels $L1 - L7$ were chosen to provide a fixed region of coverage for the lower divertor and X-point region:

$$D = \{P_j \text{ for } j \in [L1, \ldots, L7]\} \tag{1}$$

For coverage of the plasma core, channels that intersect near the magnetic axis are chosen. This is done by calculating the vertical position Z_j at which each channel j intersects the vertical $R = R_{mag}$ and choosing channels for which $|Z_j - Z_{mag}|$ is below some threshold, where (R_{mag}, Z_{mag}) is the position of the magnetic axis as shown in Fig. 2. The intersection point Z_j of each channel j is then

$$Z_j = Z_{dj} + (R_{mag} - R_{dj}) \tan \theta_j \tag{2}$$

where (R_{dj}, Z_{dj}) is the j^{th} detector position and θ_j is its orientation (measured counter-clockwise from the positive R-axis). The 'core' bin C is then defined as

$$C = \{P_j \text{ for which } |Z_j - Z_{mag}| < 0.06 \times L\} \tag{3}$$

where $L = 3\ m$ is the vertical length scale of the DIII-D plasma cross section. The 6% threshold was chosen to provide a robust margin based on analyzing several example discharges and accounting for typical variation in Z_{mag}, so this threshold may need to be changed for other devices. Since R_{mag} and Z_{mag} both vary during operation, their values must be obtained from a concurrent EFIT and the number of channels in C will fluctuate.

With the C and D bins chosen to cover specific spatial locations, we can now calculate the two P_{rad} peaking factors by taking ratios of channel measurements P_j using these bins. The 'core vs all' (CVA) peaking factor is given by

$$\boxed{P_{rad}\ CVA = \frac{mean(C)}{mean(ALL\ P_j\ for\ j \notin D)}} \tag{4}$$

while the 'divertor vs. all' (XDIV) peaking factor is

$$\boxed{P_{rad}\ XDIV = \frac{mean(D)}{mean(ALL\ P_j\ for\ j \notin C)}} \tag{5}$$

Since the denominators in equations 4 and 5 exclude channels from the alternate bin of interest, evolution of P_{rad} in the core and divertor regions is decoupled. This can in principle allow detection of events like MARFEs [23] and impurity accumulations, which are marked by movement of the radiated power distribution into and out of these regions.

II.B. T_e and n_e Peaking Factor Calculations

The density and temperature peaking factors rely on measurements from the Thomson scattering diagnostic [24] on DIII-D, which uses three separate laser systems to measure T_e and n_e at specific locations along the laser path. Two of these systems are used for our peaking factor calculations - one which covers the plasma core and another covering the plasma edge (see Fig. 3).

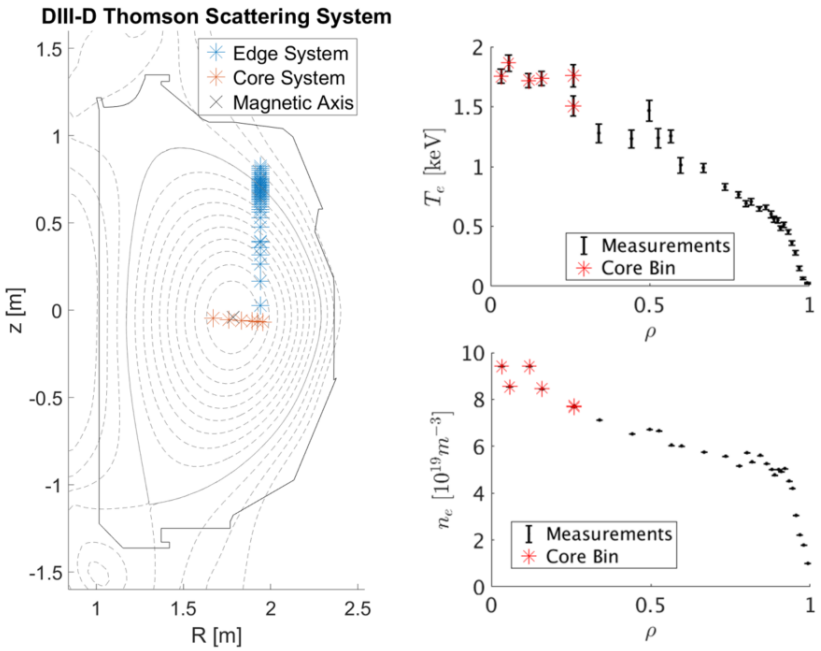

Fig. 3. (Left) geometry of Thomson scattering systems on DIII-D used for this study; (Right) typical density and temperature profiles on DIII-D, with core bin channels in red.

For each measurement location (R_j, Z_j), we obtain a value ρ_j representing the normalized

effective radius. The effective radius ρ is related to the toroidal flux ϕ via

$$\phi = B_{T0}\pi\rho^2 \tag{6}$$

where B_{T0} is the toroidal field on axis, and the normalized effective radius is simply $\rho_j = \rho/\rho_{bdry}$, where ρ_{bdry} is the value of ρ at the last closed flux surface. The core bin is then defined as

$$C = \{j|\rho_j < 0.3\} \tag{7}$$

which essentially includes all channels within 30% of the effective radius of the plasma boundary. With this in mind, the two 'core vs all' peaking factors for temperature and density are given by

$$\boxed{Te_{pf} = \frac{mean(T_j|j \in C)}{T_{avg}} \quad \text{and} \quad ne_{pf} = \frac{mean(n_j|j \in C)}{n_{avg}}} \tag{8}$$

where the denominators T_{avg} and n_{avg} represent the mean values of electron temperature and density for all channels across the entire profile. Since there were no major changes to the Thomson diagnostic during the campaigns from which our dataset was drawn, these average values are not significantly affected by large numbers of missing channels. However, the averages should be defined using a more robust method when a larger dataset is used, taking into account changes in channels available due to diagnostic upgrades or system failures. This may include incorporating a correction factor for missing channels, or defining a volume average by interpolating onto a fixed set of virtual channels.

III. DATA-DRIVEN ALGORITHMS FOR DISRUPTION PREDICTION

Being well-diagnosed experiments, laboratory plasmas are ideal testbeds for data-driven applications since many years of historical data are available for each of the existing operating fusion devices. A lot of effort has been devoted to develop reliable databases for disruption prediction [25], that usually contain information regarding many zero-dimensional features relevant for disruptive dynamics. Previous publications documented in Section I have shown how we can leverage such data abundance to develop predictive algorithms for disruption classification.

The Disruption Prediction via Random Forest (DPRF) algorithm is a supervised binary

classifier currently embedded in both DIII-D and EAST PCS. DPRF is a portable tool, so far tested on DIII-D, C-Mod, and EAST, and it was shown [12] how by using a shot-by-shot framework for simulating alarms in PCS it is possible to unravel dramatic differences in performance for disruption prediction on different tokamaks.

The Generative Topographic Mapping (GTM) is an unsupervised manifold algorithm that can trace a mapping from a latent, low-dimensional space onto the high-dimensional input space, by preserving the topology of the latter. The GTM tool has been applied to JET data to map operational boundaries for disruption prediction and to analyze the trajectory of the plasma discharge in the parameter space of interest [10].

Building on the previous work, we incorporate the peaking factor information in DPRF and verify that such engineered features are useful markers of specific disruption precursors, such as impurity accumulation events. Indeed, Figure 1 shows one example of the evolution of the 4 peaking factors on DIII-D and how their behavior closely tracks the disruption dynamics.
In the following we report on the univariate analysis of the plasma signals used for such application, by also detailing the dataset adopted. In addition, we will describe how we are gradually moving away from a fixed time threshold definition for the transition from a non-disruptive to a disruptive operational space, when describing discharges that eventually disrupt.

III.A. Univariate analysis on the input features

The DIII-D dataset used for this study was crafted from a set of thousands of discharges from the 2015-2016 campaigns, including both disruptive and non-disruptive discharges. Intentional disruptions and those caused by hardware failures and prematurely forced rampdowns were discarded, but all remaining disruptions were included regardless of the type. Additionally, only lower single null plasmas were used in order to consistently compare the P_{rad} peaking factors to those on JET. After applying these restrictions, the dataset consists of the flattop portions of 1293 shots, of which 310 were disruptions occurring during the flattop phase of the plasma current.
The JET database, as reported in [10], has been built from the first ITER-like wall (ILW) experimental campaigns (2011-2013), and consists of 114 non-disruptive discharges and 132 disrupted ones.

Conversely to what has been adopted in previous work from the authors [12,25], we illustrate here a different scheme for classifying time records in the dataset. Individual time samples from each discharge are labeled as either 'non-disruptive' or 'unstable'. All 'non-disruptive' time samples belong to non-disruptive discharges, whereas all 'unstable' time samples are taken from disruptive shots and have time $t > t_{unstable}$, where $t_{unstable}$ marks the time at which the event chain leading to disruption begins.

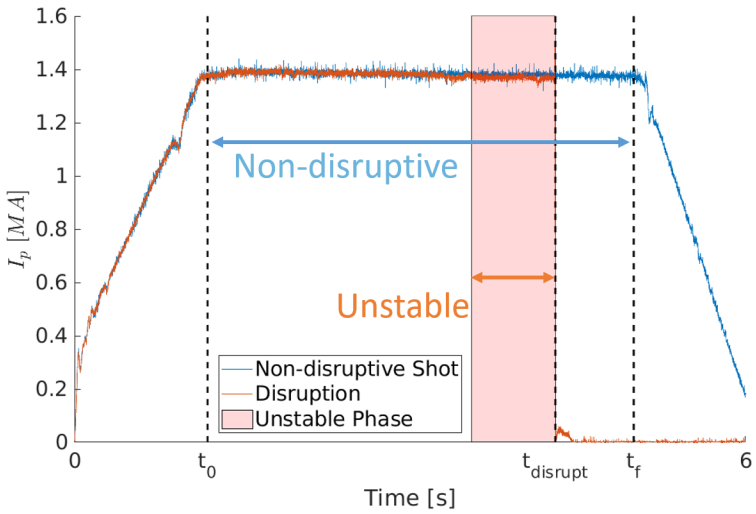

Fig. 4. Time slices in DIII-D dataset are classified based on the type of discharge they are from and (if disruptive) when they occur; both classes used in this study, i.e. 'non-disruptive' and 'unstable', are shown. t_0 and t_f define the beginning and the end of the plasma current flattop.

$t_{unstable}$ is not a fixed threshold in time as a function of the disruption event, but rather changes from shot to shot depending on the initiating cause of the disruption itself. The identification of $t_{unstable}$, i.e. the time of the first observable disruption precursor, was conducted for each disrupted discharge using a manual analysis. Generally, several attempts can be found in literature to define the beginning of the unstable phase in disruptive discharges through automatic procedures [9]. For example, in [26] the authors test several measures of similarity between probability density functions to statistically define the divergence of the unstable phase in disruptive discharges with respect to the non-disruptive parameter space. The manual identification of $t_{unstable}$ is based on the analysis conducted by Pau et al. [10,18], using the standardized precursor event descriptions first reported in a survey of disruption causes at JET [27]. As an example, we show in Figure 5 a

particular chain of events for DIII-D discharge 161238 that begins with a large drop in auxiliary power. This is followed by the start of an ELM-free H-mode phase in which impurities accumulate in the core, increasing the core radiation. Plasma performance drops significantly afterward, as an H-L back transition follows before the shot ultimately disrupts when a locked mode occurs. For this discharge, $t_{unstable}$ is thus defined as the time of the auxiliary power step-down, since this is the first event in the relevant chain of disruption precursors.

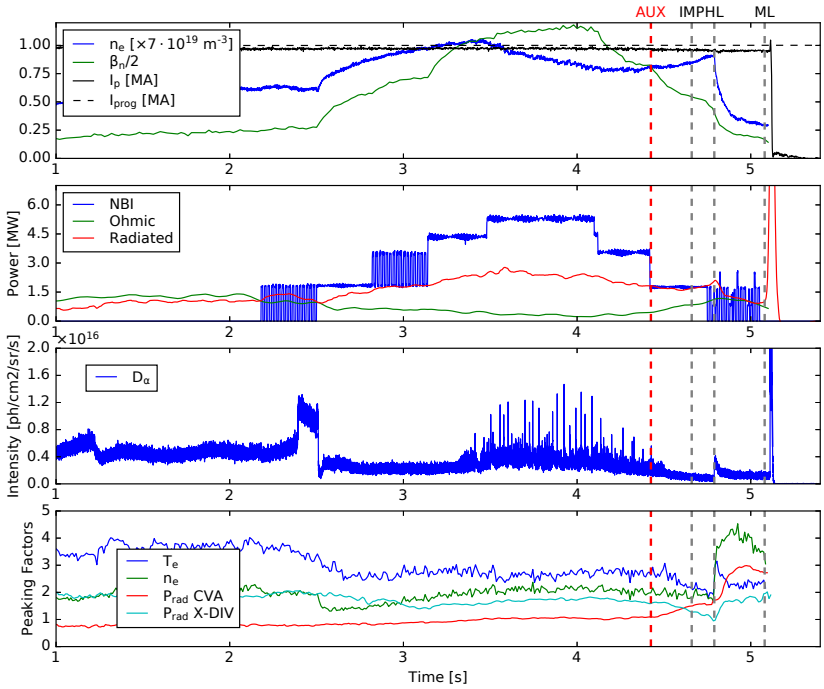

Fig. 5. Chain of disruption precursor events for DIII-D discharge #161238. The first event in the chain, marking $t_{unstable}$, is shown in red. The two- and three-letter event codes from [27] are used to identify each disruption precursor: namely, a drop in the auxiliary power (AUX), an influx of impurity (IMP) together with an H-to-L back-transition (HL), and finally a mode lock (ML).

In current analyses, all other time slices from disruptive discharges occurring in a stable phase at times prior to the beginning of the disruptive chain of events, i.e. with $t < t_{unstable}$, are excluded from the training datasets. This also allows a consistent comparison with the previous work on JET data [10], when investigating the univariate distributions of the input features used to train the algorithms presented in Section IV.

To support the hypothesis of an unstable phase in disruptive discharges, quite distinguishable

from the non disruptive flattop phase, we report in Figure 6 and 7 the univariate analysis on two of the most significant features used in the applications reported in this manuscript. Furthermore, we compare the behavior of such features, i.e. the peaking factor for the electron temperature and the Greenwald density fraction, for DIII-D and JET data.

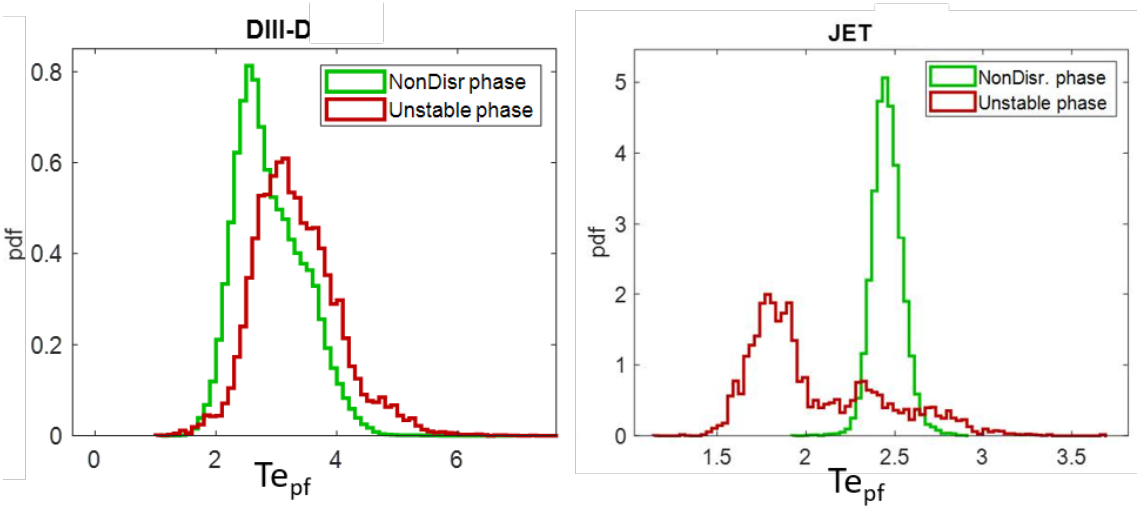

Fig. 6. Histograms of the T_e peaking factor (Left) for DIII-D and (Right) for JET data. In green the non-disruptive flattop data and in red the unstable data from disruptive shots. The different trends correlate with different statistically dominant precursors, i.e. core impurity accumulation in JET-ILW versus edge cooling mechanisms in DIII-D.

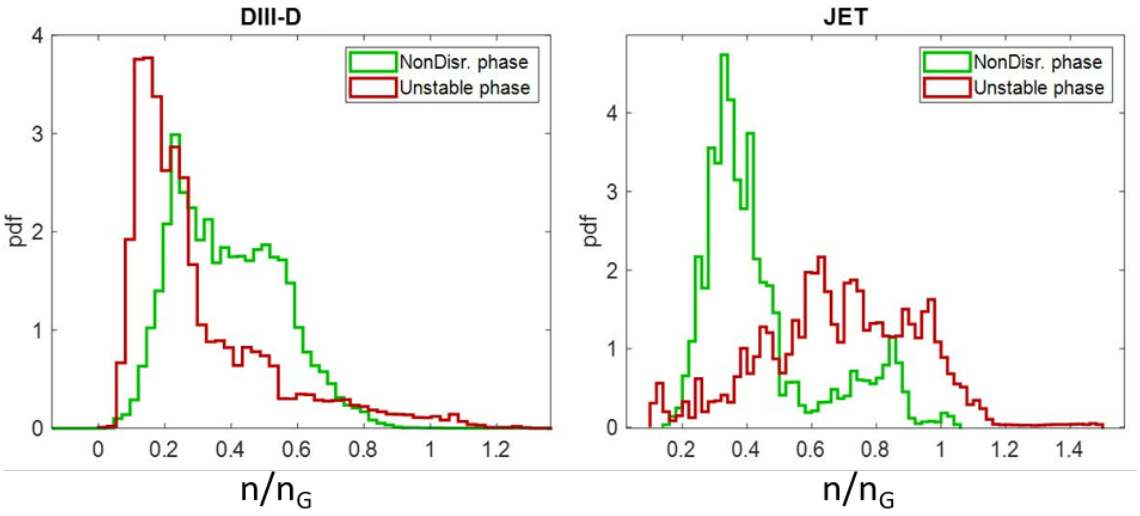

Fig. 7. Histograms of the Greenwald density fraction (Left) for DIII-D and (Right) for JET data. In green the non-disruptive flattop data and in red the unstable data from disruptive shots.

Even if the distributions for the two classes tend to be relatively well separated, it is possible to notice that the separation trend is exactly the opposite when comparing JET with DIII-D data. This highlights the very different dynamics in terms of the dominant precursors in the disruptive chain of events for the two devices: one, with its ITER-like wall (Beryllium for the main chamber wall and Tungsten for the divertor), is more prone to impurity accumulation events and therefore hollow or less peaked T_e profiles [18, 28], and the other one (with a vessel fully armored with graphite tiles) is more subject to edge cooling mechanisms, hence higher T_e peaking factors, correlated with dangerous MHD instabilities [29]. DIII-D is indeed characterized by a high error field and low-density operations (statistically frequent as shown by the Greenwald density fraction histogram in Fig. 7) allow the error field to penetrate more easily, causing locked islands that in turn flatten the temperature profile, thus cooling the peripheral regions of the plasma (see Figure 6(d) in [29]).

The probability density functions shown in Figures 6 and 7 correlate with statistically-dominant precursors that are different for the two devices, and also very dependent on the experimental dataset chosen for this application: disruptions driven by core impurity accumulation dominate JET-ILW data while MHD instabilities cause most of the analyzed DIII-D disruptions. Nevertheless, the refined definition of $t_{unstable}$, which is assigned to different plasma discharges according to the initial precursor in the disruptive chain of events, is capable to separate (in JET better than on DIII-D) the unstable scenarios from non-disruptive data. Physics-based labeling of samples for data-driven applications is extremely important to develop more robust algorithms, capable of discovering in a high-dimensional features space the boundaries between unstable regions, while successfully mapping the non-linear relationships among all the input features.

IV. PRELIMINARY ANALYSIS ON DIII-D AND JET

An efficient way to explore the predictive capability of the peaking factor metrics is to analyze them along with other relevant physics-based indicators using machine learning algorithms, which can facilitate the identification of specific recurring patterns in high-dimensional data spaces. Two disruption prediction algorithms developed by the authors, the Disruption Prediction with Random Forests (DPRF) algorithm [12] and the Generative Topographical Map (GTM) [18], are used here in an attempt to show commonalities and differences between disruption precursors on both JET

and DIII-D. In the following, we will show a few representative cases.

TABLE I
List of zero-dimensional signals used as input features in DPRF. The signal's description is given together with the associated name of the variable as it appears in this work. The last column reports the data source. All data used for the applications reported in this manuscript comes from offline data sources.

Signal description	Name	Source
Plasma current error fraction	$(I_p - I_{prog})/I_{prog}$	Rogowski Coil
Poloidal beta	β_p	EFIT
Greenwald density fraction	n_e/n_G	Interferometer and EFIT
Safety factor at 95% of minor radius	q_{95}	EFIT
Plasma internal inductance	l_i	EFIT
Radiated power fraction	$P_{frac} = P_{rad}/P_{input}$	Bolometer and Heating System
Locked mode proxy	LM	Magnetics
T_e peaking factor	Te_{pf}	Thomson Scattering
n_e peaking factor	ne_{pf}	Thomson Scattering
P_{rad} peaking factor - Core	$P_{rad}\ CVA$	Bolometer
P_{rad} peaking factor - Divertor	$P_{rad}\ XDIV$	Bolometer

In the first example of interest, DPRF is trained only on DIII-D discharges using 11 dimensionless or normalized signals that are relevant to disruption precursors, including the four peaking factors introduced in Section II. These signals, or input features, are reported in Table I. To attempt to test the portability of the algorithm and of the engineered features, DPRF is then applied to JET discharge 82657 (see Figure 8), which disrupted due to an impurity accumulation event. Note that the algorithm's output, or *disruptivity*, begins to rise substantially above the baseline value as the peaking factor marking the radiated power in the core begins to increase, as shown in the bottom panel of Fig. 8. The disruptivity is to be intended as the probability of class membership (unstable vs non-disruptive phase) and it is a dimensionless value that ranges between 0 and 1.

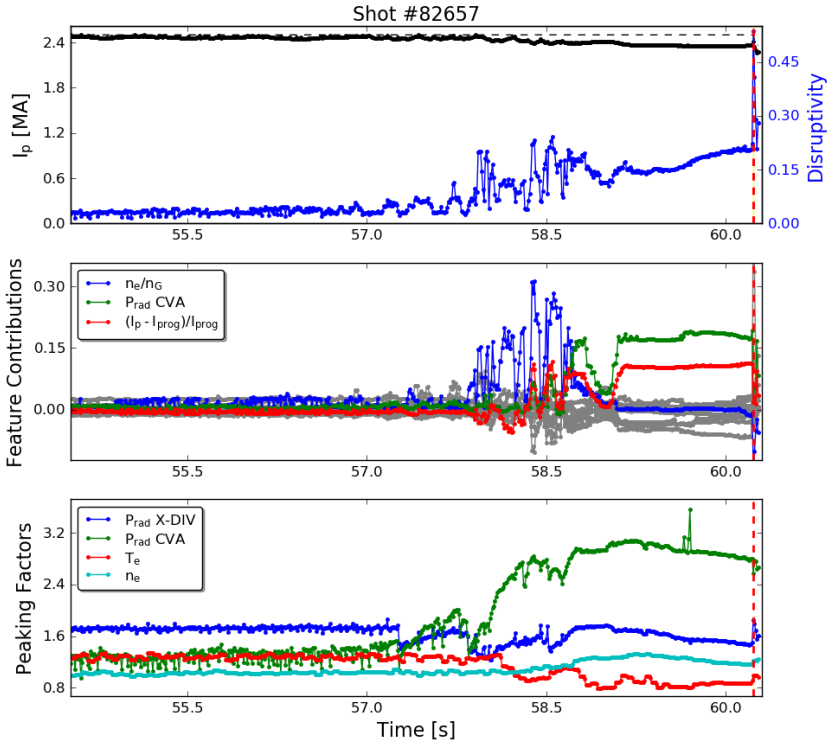

Fig. 8. JET discharge 82657 is analyzed by DPRF. Top panel shows the plasma current (black) together with the programmed I_p (dashed black) and the disruptivity prediction. DPRF is trained only using dimensionless or normalized DIII-D data as reported in Table I. Middle panel shows the feature contributions, while the peaking factor from temperature, density, and radiation are shown in the bottom panel.

An advantage of using DPRF is that its predictive output can be expressed as a sum of feature contributions [13] with the intent to reveal the extent to which each individual input feature drives the prediction. Examining these in the second panel of Fig. 8, we see that the top contributing feature during the last ∼2 s of the discharge is the $P_{rad}\ CVA$ peaking factor, indicating that DPRF is responding to an impurity accumulation in the core with a higher disruption probability. Only the most relevant contributing features are highlighted in color, while all the others are reported in grey.

In the second example, we use a slightly different set of signals (see Table II) to match the set used to train on JET-ILW discharges [10], to project a DIII-D discharge onto the latent space (i.e. lower-dimensional data-manifold) provided by the GTM algorithm.

TABLE II
List of zero-dimensional signals used as input features in the GTM. The signal's description is given together with the associated name of the variable as it appears in this work. The last column reports the data source. All data used for the applications reported in this manuscript comes from offline data sources.

Signal description	Name	Source
Safety factor at magnetic axis	q_{AX}	EFIT
Plasma internal inductance	L_i	EFIT
Radiated power fraction	P_{frac}	Bolometer and Heating System
Te peaking factor	Te_{pf}	Thomson Scattering
ne peaking factor	Ne_{pf}	Thomson Scattering
Prad peaking factor - Core	Rad_{pf-CVA}	Bolometer
Prad peaking factor - Divertor	$Rad_{pf-XDIV}$	Bolometer

Figure 9 shows the behavior of DIII-D discharge 161238, which disrupts due to three subsequent influxes of impurities: (1) at ∼2.5 s there is a transition to H-mode; analyzing the discharge's behavior on the latent space's map, it is possible to notice that the trajectory is initially evolving close to the boundary (upper left corner) due to peaked temperature and current density profiles (reported in the multi-panel plot on the right). After that, the discharge undergoes a relatively long, and stable phase away from the boundary (∼2s). Then, after the neutral beams power steps down at ∼4.5 s (not shown in figure), (2) an ELM-free H-mode phase initiates, causing an influx of impurities to flow in the plasma core, as indicated by the complementary behavior of the radiation peaking factors, shown in the set of panels on the right in Figure 9. The radiation leaves the X-point location and begins moving up toward the core plasma, with a final impurity accumulation killing definitely the discharge at 5 seconds.

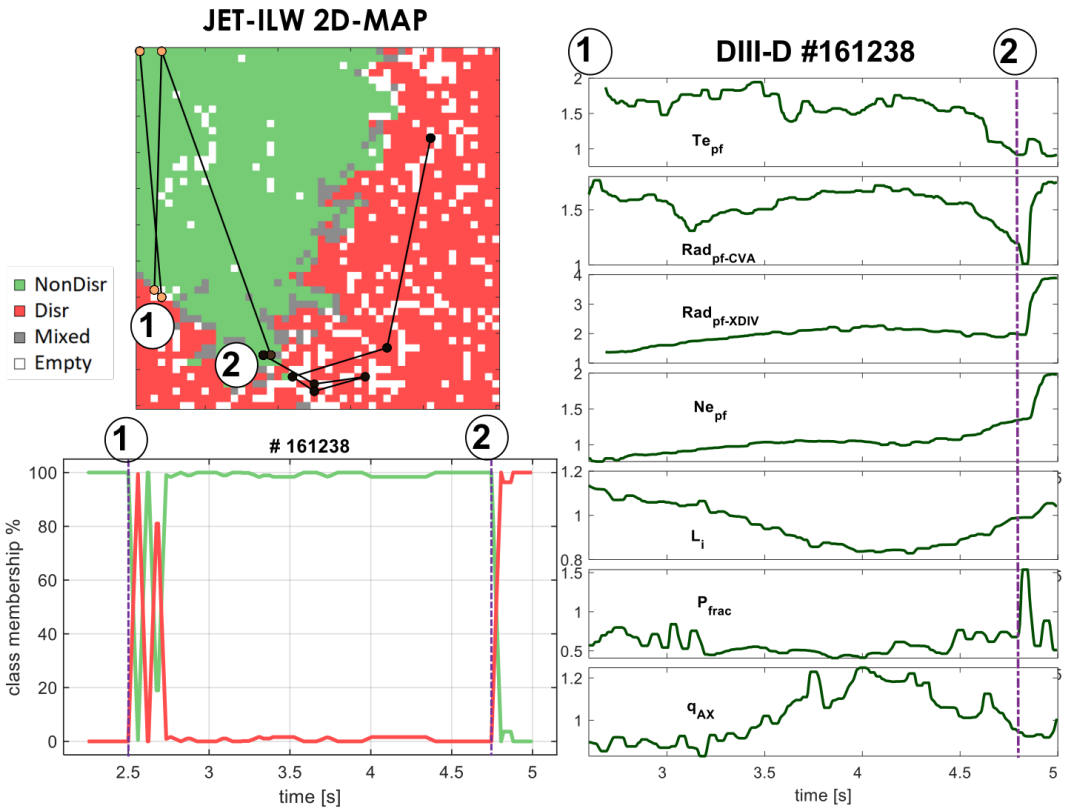

Fig. 9. 3-element figure showing (top left) the GTM latent space trained on JET-ILW data [10] with the DIII-D discharge trajectory overlaid in black with dots up to the disruption, (right) 0-D profile indicators for DIII-D discharge 161238, and (bottom left) class membership function for the same DIII-D discharge. On the right, the time traces of the input features used for the GTM, Table II describes the signals in detail.

V. DISCUSSION AND CONCLUSIONS

The preliminary analysis reported in this paper aims to lay out the groundwork for successfully developing data-driven cross-device predictors, by also preserving their interpretability. The inclusion of physics-based indicators tracking specific disruption precursors in a more uniform and device-independent framework, allows more reliable domain adaptation and knowledge extraction across very different tokamaks. The univariate analysis discussed in Section III.A illustrates how the dominant causes characterising the analysed JET-ILW and DIII-D disruption databases are statistically not the same. Despite the different frequency of occurrence of dominant precursors in the two devices, it is still possible to find a comparable ensemble of physics mechanisms leading

to disruptions, where the analogous causes can be described similarly in a unified framework of physics-based indicators. The different *statistical picture* for disruptions in JET-ILW and DIII-D is due to several factors, such as different plasma-facing component materials and device geometries, different control schemes and experimental programs. Nevertheless, we have shown how the reduction from 1D/2D profile information to 0D physics-based indicators can be used to describe specific disruption precursors that are extremely relevant also for ITER, such as impurity accumulation events, on the two very different tokamaks. In particular, through two representative examples reported in Section IV, we have shown how data-driven algorithms optimized for best performances on one device's data can still be used (with no re-training) to explain disruptive behavior on another, very different device, as long as the destabilizing precursors are diagnosed using a common definition. This is an encouraging result in view of more extended study to validate cross-device transferability of the data-driven predictors: for machines with different dominant disruptive chain of events, it may not be possible to directly transfer machine learning classifiers, even with physics-based features, without the use of at least some data from the new machine [15] (or possibly simulation data to capture the projected physics). With that being said, the inclusion of engineered profile-based features in data-driven models reveals to improve their interpretability, by tracking additional disruptive precursors, and is well suited for real-time applicability. The PCS for current and future devices will benefit from predictive models that are capable to inform on the causes of impending disruptions so that avoidance schemes can be identified. The computation of the peaking factors from 1D/2D profiles is very simple, and they are currently calculated in real-time for both JET and DIII-D: in particular, the computation of the peaking factors and the latent space on JET PCS occurs in approximately 200 microseconds. On DIII-D, recent upgrades to DPRF have included the real-time computation of the peaking factors as well [30]: to calculate in real-time the algorithm's disruptivity output and the feature contributions takes an average of 200 microseconds, similarly to the JET case.

The analyses reported in this paper leverage a more classical feature engineering approach to properly identify and simply describe the disruption mechanisms on different tokamaks through the same features and/or combination of features. The manual identification of $t_{unstable}$ to mark the beginning of loss of control that eventually leads to a disruption allows machine learning to be used

for *event-based* predictions rather than to result in a generic black-box application. Such paradigm, among other things, is compatible with the most advanced control techniques currently being explored and developed for ITER. As described in [31], the supervisory layer is tokamak-agnostic and designed as a task-based approach, with an actuator manager taking high-level decisions for handling different control tasks, such as those associated to off-normal events defined in the framework of disruption avoidance.

Possible future research directions include work to combine unsupervised (GTM) and supervised (DPRF) algorithms for disruption prediction, and leverage the existing abundance of data coming from many other existing tokamaks, such as Alcator C-Mod, TCV, EAST and KSTAR, to develop a more *global* framework that would encompass JET and DIII-D, for extrapolation to ITER.

ACKNOWLEDGMENTS

The authors would like to thank T. Odstrcil for the tomographic reconstructions of the bolometer emissivity of DIII-D discharge 175697. This work was supported by the U.S. Department of Energy under DE-FC02-04ER54698 and DE-SC0014264. Part of the data analysis reported in this paper was performed using the OMFIT integrated modelling framework [32]. DIII-D data shown in this paper can be obtained in digital format by following the links at https://fusion.gat.com/global/D3D_DMP. This work has also been carried out within the framework of the EUROfusion Consortium and received funding from the EURATOM research and training programme 20142018 and 20192020 under grant agreement No. 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission. Additionally, the SPC authors are supported in part by the Swiss National Foundation.

DISCLAIMER

REFERENCES

[1] Ikeda K 2009 *Nuclear Fusion* **50** 014002 URL https://doi.org/10.1088%2F0029-5515%2F50%2F1%2F014002

[2] Brunner D 2019 *Bulletin of the American Physical Society*

[3] Humphreys D, Ferron J, Garofalo A, Hyatt A, Jernigan T, Johnson R, La Haye R, Leuer J, Okabayashi M, Penaflor B, Scoville J, Strait E, Walker M and Whyte D 2003 *Fusion Eng. Des.* **66-68** 663–667 ISSN 09203796 URL https://linkinghub.elsevier.com/retrieve/pii/S0920379603003223

[4] Strait E J, Barr J L, Baruzzo M, Berkery J, Buttery R J, de Vries P C, Eidietis N W, Granetz R, Hanson J M, Holcomb C T, Humphreys D A, Kim J, Kolemen E, Kong M, Lanctot M J, Lehnen M, Lerche E, Logan N, Maraschek M, Okabayashi M, Park J K, Pau A, Pautasso G, Poli F M, Rea C, Sabbagh S A, Sauter O, Schuster E, Sheikh U A, Sozzi C, Turco F, Turnbull A D, Wang Z, Wehner W P and Zeng L 2019 *Nucl. Fusion* 0–18 ISSN 0029-5515 URL http://iopscience.iop.org/article/10.1088/1741-4326/ab15de

[5] Walker M L, Welander A, Humphreys D, Ambrosino G, De Tommasi G, Bremond S, De Vries P, Snipes J, Rimini F and Treutterer W 2019 *Fusion Eng. Des.* **146** 1853–1857 ISSN 09203796 URL https://doi.org/10.1016/j.fusengdes.2019.03.050

[6] De Vries P, Johnson M and Segui I 2009 *Nucl. Fusion* **49** 055011 ISSN 0029-5515 URL http://stacks.iop.org/0029-5515/49/i=5/a=055011?key=crossref.8d24467ee731fdf43d48442fbabb5f86

[7] Cannas B, Cau F, Fanni A, Sonato P, Zedda M and Contributors J E 2006 *Nucl. Fusion* **46** 699–708 ISSN 0029-5515 URL http://stacks.iop.org/0029-5515/46/i=7/a=002?key=crossref.b5d25ce338667207f051f507afd6d691

[8] Rattá G, Vega J, Murari A, Vagliasindi G, Johnson M and de Vries P 2010 *Nucl. Fusion* **50** 025005 ISSN 0029-5515 URL http://stacks.iop.org/0029-5515/50/i=2/a=025005?key=crossref.6240d82066621bd84181169f23e7fdb9

[9] Berkery J W, Sabbagh S A, Bell R E, Gerhardt S P and LeBlanc B P 2017 *Phys. Plasmas* **24** 056103 ISSN 1070-664X URL http://aip.scitation.org/doi/10.1063/1.4977464

[10] Pau A, Fanni A, Carcangiu S, Cannas B, Sias G, Murari A and Rimini F 2019 *Nucl. Fusion* **59** 106017 ISSN 0029-5515

[11] Vega J, Dormido-Canto S, López J M, Murari A, Ramírez J M, Moreno R, Ruiz M, Alves D and Felton R 2013 *Fusion Eng. Des.* **88** 1228–1231 ISSN 09203796 URL http://linkinghub.elsevier.com/retrieve/pii/S0920379613002974

[12] Montes K J, Rea C, Granetz R, Tinguely R A, Eidietis N W, Meneghini O, Chen D, Shen B, Xiao B, Erickson K and Boyer M D 2019 *Nuclear Fusion* URL http://iopscience.iop.org/10.1088/1741-4326/ab1df4

[13] Rea C, Montes K, Erickson K, Granetz R and Tinguely R 2019 *Nucl. Fusion* **59** 096016 ISSN 0029-5515 URL https://iopscience.iop.org/article/10.1088/1741-4326/ab28bf

[14] Ferreira D R, Carvalho P J and Fernandes H 2020 *IEEE Transactions on Plasma Science* **48** 36–45 URL https://doi.org/10.1109/TPS.2019.2947304

[15] Kates-Harbeck J, Svyatkovskiy A and Tang W 2019 *Nature* **568** 526–531 ISSN 0028-0836 URL http://dx.doi.org/10.1038/s41586-019-1116-4http://www.nature.com/articles/s41586-019-1116-4

[16] Montavon G, Samek W and Müller K R 2018 *Digital Signal Processing* **73** 1–15 ISSN 10512004 (*Preprint* 1706.07979) URL https://doi.org/10.1016/j.dsp.2017.10.011https://linkinghub.elsevier.com/retrieve/pii/S1051200417302385

[17] Sundararajan M, Taly A and Yan Q 2017 *34th International Conference on Machine Learning, ICML 2017* **7** 5109–5118 (*Preprint* 1703.01365) URL http://arxiv.org/abs/1703.01365

[18] Pau A, Fanni A, Cannas B, Carcangiu S, Pisano G, Sias G, Sparapani P, Baruzzo M, Murari A, Rimini F, Tsalas M and De Vries P C 2018 *IEEE Transactions on Plasma Science* **46** 2691–2698 ISSN 0093-3813 URL https://doi.org/10.1109/TPS.2018.2841394

[19] De Vries P, Pautasso G, Nardon E, Cahyna P, Gerasimov S, Havlicek J, Hender T, Huijsmans G, Lehnen M, Maraschek M, Markovič T and Snipes J 2016 *Nuclear Fusion* **56**

026007 ISSN 0029-5515 URL http://stacks.iop.org/0029-5515/56/i=2/a=026007?key=crossref.6719c8e40c5628a9c3ef00ee78fcfc84

[20] Sweeney R, Choi W, La Haye R, Mao S, Olofsson K and Volpe F 2017 *Nuclear Fusion* **57** 016019 ISSN 0029-5515 (*Preprint* 1606.04183) URL https://iopscience.iop.org/article/10.1088/0029-5515/57/1/016019

[21] Sozzi C, Alessi E, Pau A, Fanni A, Cannas B, Carcangiu S, Sias G and Sparapani P 2018 Early Identification of Disruption Paths for Prevention and Avoidance *27th IAEA Fusion Energy Conference IAEA CN-258* pp EX/P1–22 URL https://conferences.iaea.org/event/151/contributions/6273/

[22] Leonard A W, Meyer W H, Geer B, Behne D M and Hill D N 1995 *Review of Scientific Instruments* **66** 1201 URL http://dx.doi.org/10.1063/1.1146006

[23] Lipschultz B 1987 *J. Nucl. Mater.* **145-147** 15–25 ISSN 00223115 URL https://linkinghub.elsevier.com/retrieve/pii/0022311587903060

[24] Ponce-Marquez D M, Bray B D, Deterly T M, Liu C and Eldon D 2010 *Review of Scientific Instruments* **81** 10D525 URL https://doi.org/10.1063/1.3495759

[25] Rea C and Granetz R S 2018 *Fusion Sci. Technol.* **74** 89–100 ISSN 1536-1055 URL https://doi.org/10.1080/15361055.2017.1407206https://www.tandfonline.com/doi/full/10.1080/15361055.2017.1407206

[26] Alessi E, Capano M G, Pau A and Sozzi C 2019 *46th EPS Conference on Plasma Physics, EPS 2019* 1–4

[27] De Vries P, Johnson M, Alper B, Buratti P, Hender T, Koslowski H, Riccardo V and Contributors J E 2011 *Nucl. Fusion* **51** 053018 URL https://doi.org/10.1088/0029-5515/51/5/053018

[28] De Vries P C, Arnoux G, Huber A, Flanagan J, Lehnen M, Riccardo V, Reux C, Jachmich S, Lowry C, Calabro G, Frigione D, Tsalas M, Hartmann N, Brezinsek S, Clever M, Douai D, Groth M, Hender T C, Hodille E, Joffrin E, Kruezi U, Matthews G F, Morris J, Neu R, Philipps V, Sergienko G and Sertoli M 2012 *Plasma Phys. Control. Fusion* **54**

124032 ISSN 0741-3335 URL http://stacks.iop.org/0741-3335/54/i=12/a=124032?key=crossref.3aa3f02570a7e916eb49bfed5aeaaa16

[29] Du X D, Shafer M W, Hu Q M, Evans T E, Strait E J, Ohdachi S and Suzuki Y 2019 *Phys. Plasmas* **26** 042505 ISSN 1070-664X URL http://aip.scitation.org/doi/10.1063/1.5085329

[30] Rea C, Montes K J, Hu W, Erickson K, Pau A, Granetz B, Barr J L, Sammuli B, Yuan Q, Chen D L, Shen B, Xiao B J, the DIII-D Team and the EAST Team 2020 Disruption Prevention via Interpretable Data-Driven Algorithms on DIII-D and EAST *IAEA FEC 2020*

[31] Vu N T, Blanken T, Felici F, Galperti C, Kong M, Maljaars E and Sauter O 2019 *Fusion Engineering and Design* **147** 111260 ISSN 0920-3796 URL http://www.sciencedirect.com/science/article/pii/S0920379619307380

[32] Meneghini O, Smith S, Lao L, Izacard O, Ren Q, Park J, Candy J, Wang Z, Luna C, Izzo V, Grierson B, Snyder P, Holland C, Penna J, Lu G, Raum P, McCubbin A, Orlov D, Belli E, Ferraro N, Prater R, Osborne T, Turnbull A and Staebler G 2015 *Nucl. Fusion* **55** 083008 ISSN 0029-5515 URL http://stacks.iop.org/0029-5515/55/i=8/a=083008?key=crossref.5f4846d96e96da6689b641740716977c

Fig. 1: Peaking factor and profile evolution during DIII-D shot 175697, with tomographic reconstructions of bolometer emissivity at 4 different times of interest [credit pyTomo - T. Odstrcil]; in the bottom three panels, ρ refers to the normalized effective radius given in Eq. 6 and the channel number is that of the lower fan as shown in Fig. 2.

Fig. 2: (Left) Bolometer arrays on DIII-D; (Right) Lower fan channel setup, with divertor channels $L1 - L7$ shown in red.

Fig. 3: (Left) geometry of Thomson scattering systems on DIII-D used for this study; (Right) typical density and temperature profiles on DIII-D, with core bin channels in red.

Fig. 6: Histograms of the T_e peaking factor (Left) for DIII-D and (Right) for JET data. In green the non-disruptive flattop data and in red the unstable data from disruptive shots. The different trends highlight different dominance in precursors, i.e. impurity accumulation in the divertor in JET-ILW versus edge cooling mechanisms in DIII-D.

Fig. 7: Histograms of the Greenwald density fraction (Left) for DIII-D and (Right) for JET data. In green the non-disruptive flattop data and in red the unstable data from disruptive shots.

Fig. 4: Time slices in DIII-D dataset are classified based on the type of discharge they are from and (if disruptive) when they occur; both classes used in this study, i.e. 'non-disruptive' and 'unstable', are shown.

Fig. 8: JET discharge 82657 is analyzed by DPRF. Top panel shows the plasma current (black) together with the programmed I_p (dashed black) and the disruptivity prediction. DPRF is trained only using dimensionless or normalized DIII-D data as reported in Table I. Middle panel shows the feature contributions, while the peaking factor from temperature, density, and radiation are shown in the bottom panel.

Fig. 9: 3-element figure showing (top left) the GTM latent space trained on JET-ILW data [10], (right) 0-D profile indicators for DIII-D discharge 161238, and (bottom left) class membership function for the same DIII-D discharge. Note that on the GTM latent space is reported the DIII-D discharge trajectory, showing how it evolves towards a final disruption.On the right, the time traces of the input features used for the GTM, Table II describes the signals in detail.

TABLE I: List of zero-dimensional signals used in DPRF. The signal's description is given together with the the associated name of the variable as it appears in this work. The last column reports the data source.

TABLE II: List of zero-dimensional signals used as input features in the GTM. The signal's description is given together with the the associated name of the variable as it appears in this work. The last column reports the data source.