

PAPER • OPEN ACCESS

Real-time disruption prediction in multi-dimensional spaces leveraging diagnostic information not available at execution time

To cite this article: J. Vega *et al* 2024 *Nucl. Fusion* **64** 046010

View the [article online](#) for updates and enhancements.

You may also like

- [Completion of JT-60SA construction and contribution to ITER](#)
Y. Kamada, E. Di Pietro, M. Hanada et al.
- [Progress of the JT-60SA project](#)
P. Barabaschi, Y. Kamada, H. Shirai et al.
- [Progress of the JT-60SA project](#)
Y. Kamada, P. Barabaschi, S. Ishida et al.

Real-time disruption prediction in multi-dimensional spaces leveraging diagnostic information not available at execution time

J. Vega^{1,*} , S. Dormido-Canto² , R. Castro¹ , J.D. Fernández² , A. Murari^{3,4} 
and JET Contributors^a

¹ Laboratorio Nacional de Fusión, CIEMAT, Madrid, Spain

² Departamento de Informática y Automática—UNED, Madrid, Spain

³ Consorzio RFX (CNR, ENEA, INFN, Università di Padova, Acciaierie Venete SpA), Corso Stati Uniti

4, 35127 Padova, Italy

⁴ Istituto per la Scienza e la Tecnologia dei Plasmi, CNR, Padova, Italy

E-mail: jesus.vega@ciemat.es

Received 1 December 2023, revised 30 January 2024

Accepted for publication 12 February 2024

Published 23 February 2024



Abstract

This article describes the use of privileged information to train supervised classifiers, applied for the first time to the prediction of disruptions in tokamaks. The objective consists of making predictions with real-time signals during the discharges (as usual) but after training the predictor also with any kind of data at training time that is not available during discharge execution. The latter kind of data is known as privileged information. Taking into account the limited number of foreseen real time signals for disruption prediction at the beginning of operation in JT-60SA, a predictor with a line integrated density signal and the mode lock signal as privileged information has been developed and tested with 1437 JET discharges. The success rate with positive warning time has been improved from 45.24% to 90.48% and the tardy detection rate has diminished from 50% to 8.33%. The use of privileged information in an adaptive way also provides a remarkable reduction of false alarms from 11.53% to 1.15%. The potential of the methodology, exemplified with data relevant to the beginning of JT-60SA operation, is absolutely general and can be applied to any combination of diagnostic signals.

Keywords: disruption prediction, privileged information, JT-60SA, SVMplus

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

^a See Mailloux *et al* 2022 (<https://doi.org/10.1088/1741-4326/ac47b4>) for JET Contributors.

* Author to whom any correspondence should be addressed.



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

1. Prediction for mitigation when missing important diagnostic information

An essential control task in next Tokamaks (such as JT-60SA, ITER or DEMO) will be the reliable real-time prediction of forthcoming disruptions. In these devices, the energy content of the plasma can be extremely high and, therefore, unmitigated disruptions could have catastrophic repercussions on the machine integrity.

Since more than a decade, as a consequence of the lack of full disruption models from first principles [1, 2], the recognition of disruptive events has been carried out by means of machine learning techniques [3]. Typically, supervised classification methods based on two classes of examples (disruptive and non-disruptive) have been developed. The different techniques (for instance, SVM [4], artificial neural networks [5], fuzzy logic [6], nearest centroid [7], random forest [8] or conformal prediction [9, 10]) split the parameter space into two zones through such a separation frontier. To obtain such frontier, a training process with examples of both classes is carried out. From a mathematical point of view, the examples are represented by feature vectors $\mathbf{x}$ in multi-dimensional parameter spaces (typically, $\mathbf{x} \in R^m$). Each one of the m dimensions can be one of the following: (a) quantities in the time domain [11], (b) quantities in the frequency domain [12, 13] or (c) quantities that result from particular signal transforms [14]. In general, parameter spaces can comprise simultaneously several of the previously mentioned features [9, 15, 16].

Nowadays, in spite on the strong ‘black-box’ character of deep learning classifiers, they generate a lot of interest to detect approaching disruptions [17–21]. The main drawback of this type of classifiers is the large amount of data that are necessary for their training. In this respect, they could be used for transfer learning purposes [22, 23], i.e. to train models with existing disruption databases (JET, Asdex Upgrade, DIII-D, KSTAR, EAST, JT-60U and others) and to apply the models to new devices (JT-60SA, ITER or DEMO). However, although this option is very promising, it still requires further research and validation.

Other approaches related to transfer learning are not based on ‘black-box’ models but on finding relations between physics quantities [24–26]. Indubitably, this is the best option to model disruptions. However, although big efforts are being carried out at present, it is a difficult task and only partial success has been achieved so far.

For next generation Tokamaks, a potential data-driven alternative is the development of adaptive classifiers from scratch [9, 16, 27–30]. This means to create data-driven models with the data produced from the devices themselves since the beginning of their operation. When prediction errors occur, the classifiers are re-trained in an adaptive way to incorporate new information in order to improve the success rate.

Once a classifier of any type has been trained (i.e. it is able to distinguish between disruptive and non-disruptive behaviours in a reliable way), it can be installed in the real-time network of a Tokamak (examples of predictors in the JET real-time network are APODIS [15], SPAD [14] and the Centroid

Method [7, 31]). As a discharge is being run, feature vectors are generated on a periodic basis and are sent to the model to be classified. It should be noted that the dimensions and features of these vectors are the same that were used to train the model. As a result of the classification, when a feature vector is located in the non-disruptive (disruptive) zone of the classifier, a non-disruptive (disruptive) behaviour is predicted. This prediction capability is the reason to call these classifiers ‘*disruption predictors*’. Of course, in the case of the predictor detecting an incoming disruption, an alarm has to be triggered for the Tokamak control system to undertake the necessary actions.

Coming back to the feature vectors and the physics quantities from which the features are defined, it is important to mention that there is a certain consensus on which datasets of signals to adopt for each type of predictors. Some of the typical signals are the mode lock, the internal inductance, the total input power, radiated power, density and radiation monitors among others [32, 33]. Probably, the most important one is the mode lock as explained in the next paragraph.

In general, disruptions take place when the plasma is pushed close to an operational limit, whose consequence is the onset of physics instabilities. These destabilizing phenomena generate a sequence of events, which eventually lead to the disruption. The chains of events are complex combinations of physics instabilities, whose last step is usually the locking of macroscopic modes to the wall [34]. It should be noted that mode locking is in direct relation to MHD instabilities [35, 36]. Therefore, the mode lock signal has been used typically as disruption precursor for mitigation purposes, as it appreciably grows close to the beginning of the current quench. This signal has shown to be an essential feature not only in single-signal predictors [7, 14] but also in multi-dimensional ones [37, 38].

Due to the prominence of the mode lock signal to predict disruptions (at least for mitigation purposes), this article analyses how to proceed and handle those situations, in which such a signal is not available. On the one hand, the signal unavailability can be related to inadequate signal conditioning due to distortions or noise. In these cases, off-line signal processing may filter undesirable components and increase the signal quality for off-line analysis. On the other hand, the mode lock amplitude simply cannot be available in real-time but being estimated by specific codes only after the discharge. This second situation is what will happen in the initial operation of JT-60SA. In the two cases just mentioned, the mode lock signal cannot be used in feedback during the discharges; however trustworthy mode lock signals can be obtained in a deferred way by means of off-line analysis.

Owing to the fact that the mode lock signal is essential for disruption prediction but under some circumstances it cannot be accessed in real-time, an important question arises: can off-line versions of the mode lock signal be used for disruption prediction? The answer is affirmative by means of the ‘*Learning Using Privileged Information*’ (LUPI) paradigm [39]. According to this conceptual approach, the objective is to generate classifiers making use of all possible available information at training time but only using the accessible information at execution time.

Therefore, if the mode lock is not available to make real-time predictions, which other signal can be used by the classifiers? Can the predictions of this potential signal be improved with the mode lock by applying the LUPi paradigm?

The detailed answer to the first question depends of course on the circumstances and the status of the other diagnostics. To fix the ideas, in the following, the discussion is particularised for the case of the JT-60SA device and the list of available real-time quantities during the first phases of its operation [40]. However, the approach is of course absolutely general. With regard to disruptions in JT-60SA, there will be a set of real-time signals linked to the current quench (for example, plasma current, loop voltage, on-axis toroidal field, diamagnetic flux and plasma radial shift) that can be used for prediction purposes as shown in [41]. Also, it is important to note that a line integral density (LID), measured by interferometry, will be available in real-time. According to [42], a line density diagnostic can be used to detect MHD instabilities. Therefore, the capabilities of LID as disruption predictor can be evaluated. It is important to point out that the simple amplitude of the LID is not valid to make predictions, because a threshold in the LID signal is not enough to identify disruptive conditions. Nevertheless, the parameter space of consecutive samples will prove to be sufficiently informative, as described in the present work.

In relation to the question about the application of LUPi to disruption prediction, it should be noted that, for instance, in the case of JT-60SA, the amplitude of the mode lock signal will be available for post-process analysis. Therefore, the offline signal can be used as privileged information at training time.

This article analyses the application of LUPi for disruption prediction with SVM, in particular, a version called SVM+. The article develops an extreme case with only two signals: a LID signal to be used at training and execution times together with a mode lock signal that is used as privileged information during the training process. As shown in [41], of course other combinations of signals could be equally effective.

On the other hand, it is important to note that, for mitigation, predictors requiring only one signal could be very advantageous particularly at the beginning of the operation of new devices, when the diagnostic capability is typically very limited. However, Vega *et al* discuss in [7] that just setting a threshold amplitude in a single-signal predictor is not reliable enough. One alternative for the use of single-signal predictors is to build a specific two-dimensional parameter space, whose components are the amplitudes of consecutive samples. In this space, sudden changes in the signals can be detected and, depending on the signal, can be connected to the presence of disruptive behaviours. This is tested in [7] with the mode lock signal normalised to the plasma current. The rates of detected disruptions and false alarms are respectively 98% and 4%. However, it should be taken into account that in the analysis proposed in the present article, the mode lock is not available in real-time. But if it is possible to estimate its value by offline analysis, the mode lock signal can be used as privileged information, as it will be shown in the following.

The rest of the paper is organised as follows. Section 2 describes the database of signals and discharges that are used. Section 3 reviews the standard SVM algorithm and section 4 shows the SVM+ algorithm, i.e. the extension of SVM to address LUPi. Section 5 presents four different training methodologies: standard and adaptive training of SVM classifiers, standard and adaptive training of SVM+ classifiers, comparing their prediction performances. Finally, section 6 is a short discussion.

2. Signals, discharges and parameter space for training/test purposes

The database to test disruption prediction with the LUPi paradigm is made up of 1437 JET discharges with the C wall in the range 65 988–73 126 (April 2006–June 2008). The selection of C wall discharges is related to the fact that JT-60SA will start its operation with graphite plasma facing components. In the database, there are 85 unintentional disruptive shots and 1352 non-disruptive ones. Only discharges with flat top plasma currents above 2 MA have been considered. In addition to this, only disruptions whose plasma current at disruption time is greater than 1.5 MA are taken into account. The disruption time is assumed to be the start of the current quench. Finally, it is important to mention that the sampling period of the signals has been 2 ms, which is the maximum temporal resolution of the JET real-time data network.

Figure 1 shows the layout of the JET interferometer diagnostic, with its 8 LID signals. In this article, a single vertical chord through the plasma centre has been chosen (LID 3 in the figure) as the most informative. As the simple LID amplitude does not allow distinguishing between disruptive and non-disruptive behaviours (discharges can have larger or smaller line density), the signal is analysed in the parameter space of consecutive samples.

The parameter space of consecutive samples was first introduced in [7]. This space maps a temporal evolution signal $x(t)$ into a two-dimensional space to relate the amplitudes of sequential samples. Given a temporal evolution signal and two consecutive samples, $x(t - \tau)$ and $x(t)$ where τ is the sampling period, these samples define one point in the space of consecutive samples in such a way that the X coordinate is $x(t - \tau)$ and the Y coordinate is $x(t)$. This allows an easy visualization to detect large jumps between consecutive samples. If the two amplitudes are similar, the point in such parameter space will appear around the diagonal. On the contrary, if $\delta = x(t) - x(t - \tau)$ is large, the resultant point will not be close to the diagonal. In other words, if a plasma quantity evolves in a smooth way ($\delta \approx 0$), which happens in a calm plasma evolution, the samples are concentrated around the diagonal and a large elongated cluster is formed. However, the presence of plasma instabilities (for example, an incoming disruption) can produce abrupt jumps between consecutive samples and, therefore, the corresponding points will be located far apart from the elongated cluster of data that show smooth evolutions (figure 2).

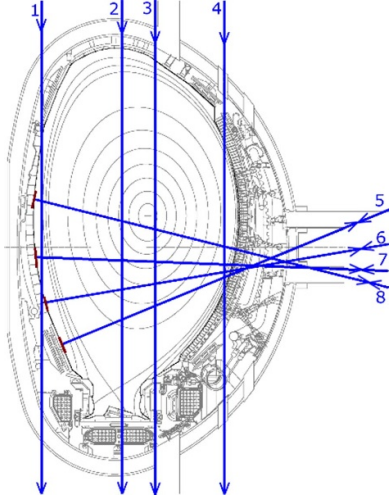


Figure 1. Layout of JET interferometer providing the line integrated density signals.

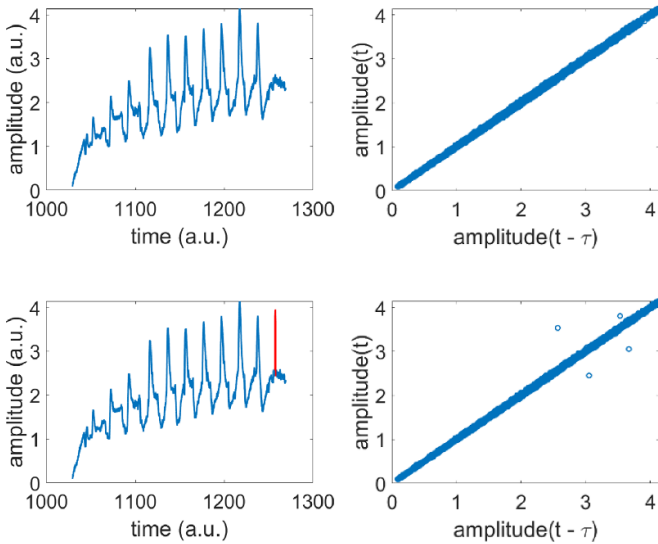


Figure 2. Arbitrary temporal evolution signal and its representation in the parameter space of consecutive samples. Top left (reference signal): temporal evolution signal that evolves in a smooth way. Top right: parameter space of consecutive samples corresponding to the reference signal. The differences of amplitudes between consecutive sample are small and the points appear around the diagonal. Bottom left: the red peak is an abrupt peak with high frequency components that has been added to the reference signal. Bottom right: four points are located far from the elongated cluster that correspond to samples of the red peak.

It should be mentioned that the use of a LID signal with the above parameter space to detect abrupt changes of amplitude is not completely reliable from a diagnostic point of view. It is indeed well known that the interferometric diagnostic can show fringe jumps and, therefore, this rapid change in the LID amplitude can be wrongly associated to an incoming disruption in the two-dimensional space of consecutive samples. This potential effect has to be dealt with at the level of the diagnostic not of the disruption predictors though. Indeed,

the next generation of devices will tend to install dispersion interferometers explicitly designed to minimise this problem. Techniques to correct for fringe jumps in real-time are also available [43].

3. Background on support vector machines for classification

SVM is a supervised classification method in multi-dimensional spaces that determines the separation frontier (decision function) between examples of two different classes. Formally, given a training set of examples $\{z_1, \dots, z_n\}$, where each $z_i \in Z = X \times Y$ is a pair $(\mathbf{x}_i, y_i)$ that is composed of the sample (or feature vector) $\mathbf{x}_i \in X$ and its corresponding class $y_i \in \{-1, +1\}$, and given a new feature vector $\mathbf{x}_{n+1}$, the objective of SVM is to estimate the class $\{-1\}$ or $\{+1\}$ to which $\mathbf{x}_{n+1}$ belongs to.

It is well-known [44] that to find the SVM decision function by means of Lagrangian formulation, a quadratic optimization problem subject to linear constraints has to be solved. Specifically, given the training set pairs $(\mathbf{x}_i, y_i)$, $i = 1, \dots, n$, one has to find the parameters $\alpha_i, i = 1, \dots, n$ that maximize the functional

$$L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j H(\mathbf{x}_i, \mathbf{x}_j) \quad (1)$$

subject to constraints

$$\begin{aligned} \sum_{i=1}^n y_i \alpha_i &= 0 \\ 0 &\leq \alpha_i \leq C, \quad i = 1, \dots, n \end{aligned}$$

where $H(\mathbf{x}_i, \mathbf{x}_j)$ is the inner product kernel.

The decision function is

$$f(\mathbf{x}) = \sum_{i=1}^n y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \quad (2)$$

and, therefore, given a new feature vector $\mathbf{x}$, the classification is provided by the output of

$$y = \text{sign}(f(\mathbf{x})) \quad (3)$$

where $\text{sign}(\cdot)$ is the sign function, i.e.

$$\text{sign}(f(\mathbf{x})) = \begin{cases} +1 & \text{if } f(\mathbf{x}) > 0 \\ -1 & \text{if } f(\mathbf{x}) < 0 \end{cases}.$$

4. Background on SVM+ for classification

As the SVM case, SVM+ is a supervised classification method in multi-dimensional spaces that determines the separation frontier (decision function) between examples of two different classes [39]. Given a training set of examples $\{z_1, \dots, z_n\}$,

where each $z_i \in Z = X \times X^* \times Y$ is a triplet $(\mathbf{x}_i, \mathbf{x}_i^*, y_i)$, $\mathbf{x}_i \in X$, $\mathbf{x}_i^* \in X^*$, $y_i \in \{-1, +1\}$ and given a feature vector $\mathbf{x}_{n+1} \in X$, SVM+ determines the class $\{-1\}$ or $\{+1\}$ to which $\mathbf{x}_{n+1}$ belongs to.

The space X is made up of vectors, whose features are available at any time (training and classifier execution). However, the feature vectors of the X^* space (which is different from the space X) are only available at training time. Due to this fact, X^* is called ‘privileged information space’.

The aim of SVM+ is to give at the training stage some additional information $\mathbf{x}^*$ about training example $\mathbf{x}$. This privileged information allows modifying the decision function that would have been obtained with the classical SVM method.

According to [39], the Lagrangian formulation for SVM+ requires to optimize a quadratic optimization problem subject to linear constraints, as it happens with classical SVM. Given the training triplets $(\mathbf{x}_i, \mathbf{x}_i^*, y_i)$, $i = 1, \dots, n$, it is necessary to find the parameters $\alpha_i, i = 1, \dots$, and $\beta_i, i = 1, \dots, n$ that are solution of the following optimization problem: maximize the functional

$$L(\alpha, \beta) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j H(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{2\gamma} \sum_{i,j=1}^n (\alpha_i + \beta_i - C)(\alpha_j + \beta_j - C) H^*(\mathbf{x}_i^*, \mathbf{x}_j^*) \quad (4)$$

subject to the constraints

$$\begin{aligned} \sum_{i=1}^n (\alpha_i + \beta_i - C) &= 0, \\ \sum_{i=1}^n y_i \alpha_i &= 0, \\ \alpha_i &\geq 0, \\ \beta_i &\geq 0. \end{aligned}$$

Here, $H(\mathbf{x}_i, \mathbf{x}_j)$ and $H^*(\mathbf{x}_i^*, \mathbf{x}_j^*)$ are kernels in X and X^* spaces that define inner products.

According to SVM+, the decision function in the X space is

$$f(\mathbf{x}) = \sum_{i=1}^n y_i \alpha_i H(\mathbf{x}_i, \mathbf{x}) + b \quad (5)$$

and the corresponding correcting function is

$$\phi(\mathbf{x}^*) = \frac{1}{\gamma} \sum_{i=1}^n (\alpha_i + \beta_i - C) H^*(\mathbf{x}_i^*, \mathbf{x}^*) + b^*.$$

Given a feature vector $\mathbf{x} \in X$ to classify, its label is obtained from

$$y = \text{sign}(f(\mathbf{x})) \quad (6)$$

where, again, $\text{sign}(\cdot)$ is the sign function.

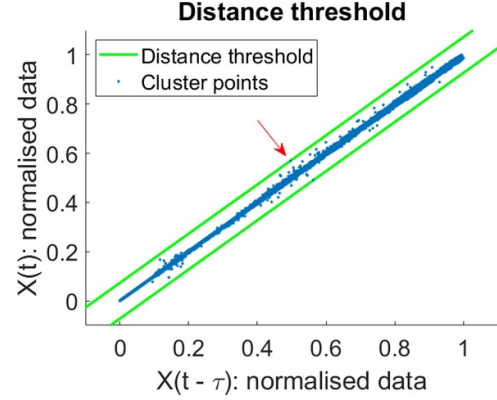


Figure 3. Elongated cluster of 42 non-disruptive discharges (605 319 points). The ellipse (in green) defined by the largest Mahalanobis distance to the cluster centre establishes the threshold distance $D_{\text{threshold}} = 52.48$ to recognise disruptive behaviours. τ is the sampling period.

5. Comparison of four different information handling and training methods

It is important to remember that the objectives of this article are (a) to test the disruption prediction capability of a LID signal; (b) to create adaptive classifiers based on the LID signal as discharges are produced and (c) to apply the LUPi paradigm.

The rationale behind the development of adaptive classifiers resides in the fact that the learning process can incorporate new knowledge as the operation evolves and new discharges are executed. Therefore, in order to mimic real life situations, the discharges will be processed in chronological way.

As mentioned, supervised classifiers based on either SVM or SVM+ will be generated. However, to create a first supervised model, at least one disruptive shot and one non-disruptive shot are required. In the dataset of 1437 discharges analysed in this article, the first 42 are non-disruptive. As mentioned the chosen feature space to develop supervised classifiers is the space of consecutive samples. In this space, the data of safe discharges tend to group in an elongated cluster, as shown in figure 3. With elongated clusters, the use of Euclidean distances is not suitable to detect points that are ‘far enough’ from the centre. To recognise an anomaly with a Euclidean distance, the corresponding point should appear outside a circumference with a radius defined by the furthest point (in Euclidean sense) to the cluster centre (iso-distance contours are circumferences). Instead, Mahalanobis distances have to be used. In this case, the iso-distance contours are ellipses and, consequently, this is the type of distance most appropriate to detect points distant from an elongated cluster [7].

The 43rd discharge is disruptive as shown in figure 4. The plot of figure 5 reports the first disruptive points in the considered feature space. Disruptive behaviours are recognised when points of such discharge are located outside the green ellipse of figure 3.

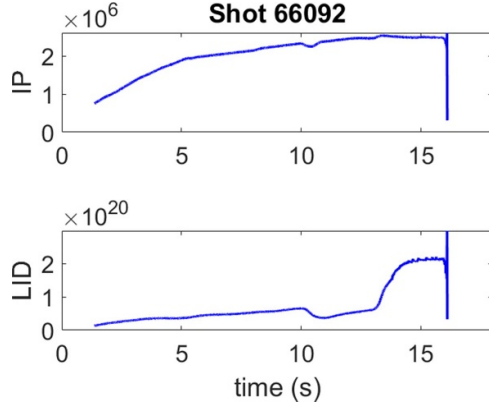


Figure 4. Time evolution of the plasma current and the LID3 signal of the first disruptive discharge, shot 66 092.

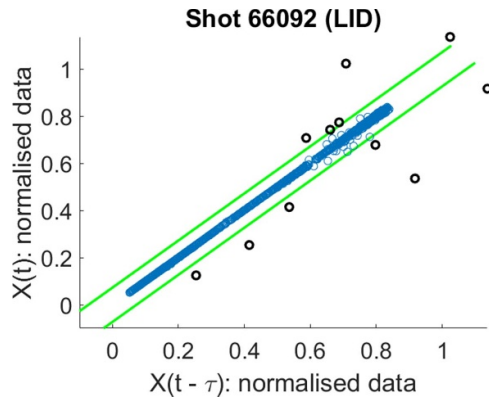


Figure 5. Elongated cluster of points in the space of consecutive samples corresponding to the LID signal of shot 66 092. Points whose Mahalanobis distance to the cluster centre is greater than the threshold defined in figure 3 are assumed to represent a disruptive behaviour. τ is the sampling period.

At this point, a supervised classifier can be trained. In the space of consecutive samples shown in figure 5, there are 11 disruptive examples and 7370 non-disruptive examples. To generate a balanced set of disruptive and non-disruptive examples, all the disruptive points are chosen to train a classifier together with 11 non-disruptive points that are selected in a random and uniform way to cover the entire elongated cluster. This set of 22 feature vectors from the LID signal define the starting dataset for all the different training methodologies presented in this article.

After having defined the disruptive and non-disruptive examples that form the first dataset to create a supervised classifier, 4 different information handling and training methods are analysed. Sections 5.1 and 5.2 show results of using the standard SVM method with only the LID signal. Sections 5.3 and 5.4 analyse the results of using the mode lock as privileged information at training time and the LID signal for training and real-time prediction.

Table 1 describes the terms and criteria implemented to quantify the results. The disruption time is defined as the beginning of the current quench. Tardy alarms are those

Table 1. Meaning of terms.

Warning time	Difference between the disruption time and the time in which an alarm is triggered
Success rate	Rate of predicted disruptions
Success rate with positive warning time	Rate of predicted disruptions before the current quench
Tardy alarm rate	Rate of predicted disruptions after the current quench
False alarm rate	Rate of non-disruptive discharges predicted as disruptive ones
Premature alarms	Alarms triggered earlier than 0.5 s before the beginning of the current quench

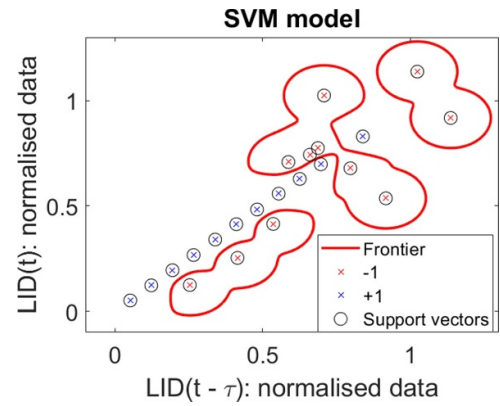


Figure 6. Red lines are the separation frontiers between disruptive (label -1) and non-disruptive (label $+1$) examples. The sampling time is $\tau = 2$ ms.

triggered after the beginning of the current quench. Alarms triggered more than 0.5 s before the beginning of the current quench are considered premature, in agreement with the results reported in [41].

5.1. Results with the standard SVM without retraining

This section develops a supervised classifier with the dataset of 22 initial examples (11 disruptive and 11 non-disruptive ones) from the LID signal. In particular, the features of each example are two-dimensional vectors, whose components are $(LID(t - \tau), LID(t))$, i.e. consecutive amplitudes of the LID signal. The sampling period τ is 2 ms that is the maximum resolution time of the JET real-time data network. Figure 6 shows how the standard SVM classifier splits the two-dimensional space of consecutive samples into disruptive (points enclosed by the red lines) and non-disruptive (rest of points) zones. The red lines are the separation frontiers that are determined by equation (2) as a result of maximizing the functional of equation (1) subject to the corresponding linear constraints. This classifier is obtained with a radial basis function (RBF) kernel [44], together with the additional condition of separating the training examples without error, i.e. there are no

Table 2. Standard SVM and no adaptive re-training. SR: Success rate. SRPWT: Success rate with positive warning time. TA: Tardy alarms. FA: False alarms. AWT: Average warning time. STD: standard deviation of the warning times.

	Rates	AWT (s)	STD (s)
SR	95.24% (80/84)	—	—
SRPWT	45.24% (38/84)	3.3	3.4
TA	50.00% (42/84)	0.006	0.004
FA	11.53% (151/1310)	—	—

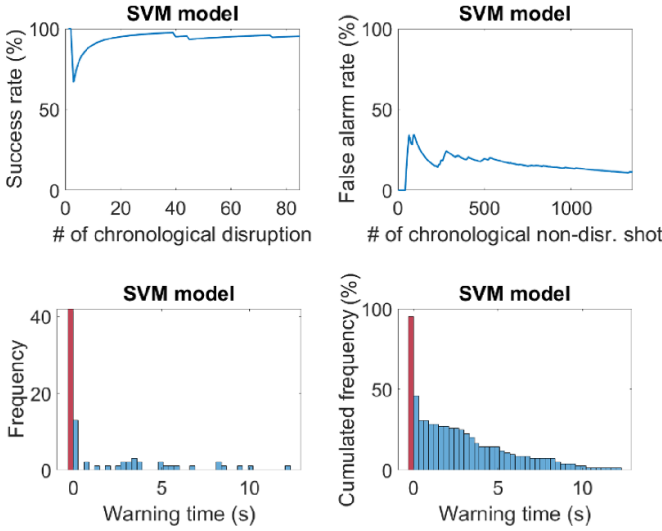


Figure 7. Top left: success rate evolution. Top right: false alarm rate evolution. Bottom left: number of correct predictions and the corresponding warning time (bin width is 300 ms). Bottom right: cumulated fraction of detected predictions at the different warning times (again, the bin width is 300 ms). The bins in red represent tardy alarms. The rate of premature alarms is about 30%.

Table 3. Standard SVM and adaptive re-training. SR: Success rate. SRPWT: Success rate with positive warning time. TA: Tardy alarms. FA: False alarms. AWT: Average warning time. STD: Standard deviation of the warning times.

	Rates	AWT (s)	STD (s)
SR	98.81% (83/84)	—	—
SRPWT	88.10% (74/84)	0.558	1.641
TA	10.71% (9/84)	0.002	0.002
FA	2.14% (28/1310)	—	—

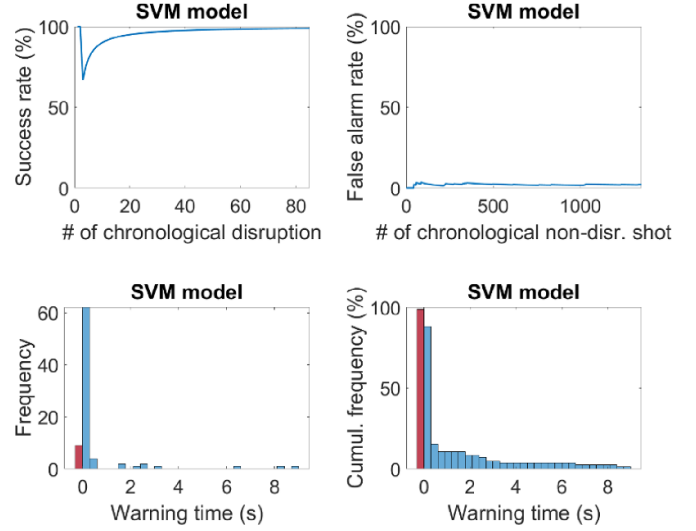


Figure 8. Top left: success rate evolution. Top right: false alarm rate evolution. Bottom left: number of correct predictions and the corresponding warning time (bin width is 300 ms). Bottom right: cumulated fraction of detected predictions at the different warning times (again, the bin width is 300 ms). The bins in red represent tardy alarms. The rate of premature alarms is about 15%.

examples on the wrong side of the decision function in the training process.

Table 2 summarises the results of applying the model of figure 6 to the dataset of JET discharges without any retraining. Each feature vector that is input to the classifier (with a time period $\tau = 2$ ms) has two components, $(LID(t - \tau), LID(t))$, the same as the feature vectors of the training set. The predictions (disruptive/non-disruptive behaviours) are carried out according to equation (3). It should be mentioned that the results are obtained assuming a real-time execution of the predictor. Also, figure 7 shows the evolution of the success rate and the false alarm rate with the chronological production of discharges together with the distribution of warning times.

It should be noted from table 2 and figure 7 the high false alarm rate in these results, although a high fraction of these false alarms come from fringe jumps in the LID signal. Moreover, it is important to mention that most of the alarms in disruptive discharges take place close to the disruption. This means that predictors with only the LID signal can be useful for mitigation purposes. Also, it is clear that there are alarms with very high warning times. These alarms have to be considered either premature alarms or false alarms as consequence of fringe jumps (in the following alarms with high warning

times will be referred to as premature alarms). On average, the warning time and its standard deviation are above 3 s. However, it is important to comment that most of the premature alarms have the origin in fringe jumps. With regard to tardy alarms, it is important to note that, on average, the alarms are triggered only 6 ms late with a standard deviation of 4 ms.

5.2. Results with the standard SVM and adaptive retraining

The starting point of this section is the first supervised classifier shown in figure 6. This means to use the LID signal for prediction purposes in the two-dimensional space of consecutive samples. Now, in order to increase the knowledge of the classifier, the aim is to re-train the model after missed or tardy alarms. This is what would happen in the real life during the experimental campaigns. The processing of discharges has been carried out by simulating real-time operation. Again, RBF kernels are used and the successive data-driven models are obtained with the extra requirement of separating the training samples without error.

Table 3 and figure 8 summarise the outcomes of SVM predictors with re-trainings. As in the case analysed in the previous section, most alarms are fired close to the disruption,

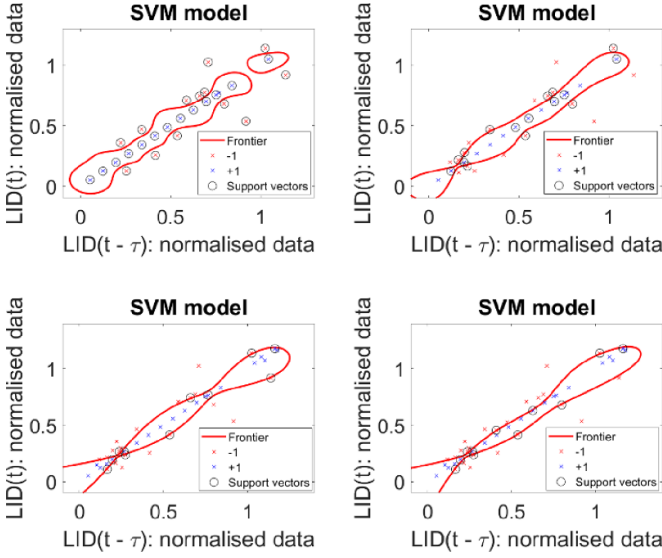


Figure 9. Resulting models after some of the re-trainings. As can be seen, the effect of the re-trainings is to change the separation frontier between disruptive and non-disruptive behaviours.

which confirms that the LID signal is valid for mitigation purposes. The adaptive training of the predictors results in an increase of the global success rate. In relation to section 5.1, this increment is a consequence of a higher success rate with positive warning time and a lesser rate of tardy detections. Moreover, the effect of adding knowledge with consecutive re-trainings has been an important reduction of false alarms.

With the approach described in this subsection, there are fewer premature alarms than in the case without retraining. But even taking into account this kind of alarms, the average warning time is less than 600 ms and the standard deviation is slightly above 1.6 s. This standard deviation means that the warning times of the premature alarms are quite widely distributed. On the other hand, the distribution of tardy warning times shows a mean value of 2 ms with a standard deviation of 2 ms.

To conclude this section, figure 9 shows some examples about the way in which the classifiers have evolved. It is important to remember that the first classifier corresponds to the one of figure 6. After 1 missed alarm and 9 tardy alarms, there were 10 re-trainings and figure 9 shows some of the resulting models. It should be noted that as the model is retrained with new points after missed or tardy alarms, the parameter space is split into two zones. Firstly, one zone around the diagonal that represents non-disruptive behaviours. Secondly, the rest of the space that determines disruptive behaviours. It is important to emphasise that in all cases, the training points are classified in a correct way.

5.3. Results with privileged information but without retraining

The present section shows the results of creating one classifier with privileged information and applying it to the database of JET discharges. At training time, there are 11 disruptive

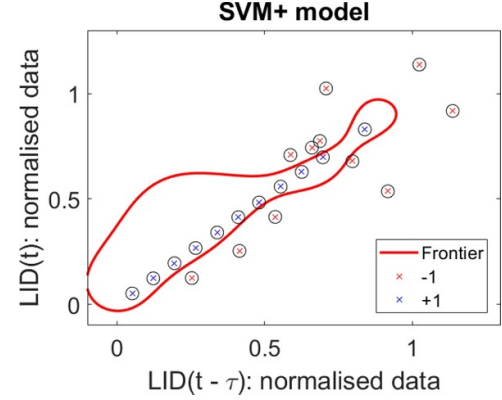


Figure 10. The red lines is the separation frontier between disruptive (label -1) and non-disruptive (label $+1$) examples. The sampling time is $\tau = 2$ ms.

Table 4. Predictor with privileged information without adaptive re-training. SR: Success rate. SRPWT: Success rate with positive warning time. TA: Tardy alarms. FA: False alarms. AWT: Average warning time. STD: Standard deviation of the warning times.

	Rates	AWT (s)	STD (s)
SR	98.81% (83/84)	—	—
SRPWT	63.10% (53/84)	1.7	2.3
TA	35.71% (30/84)	0.005	0.005
FA	5.04% (66/1310)	—	—

examples and 11 non-disruptive ones. They are the same used in section 5.1 but now, the feature vectors have 4 components: $(LID(t - \tau), LID(t), ML(t - \tau), ML(t))$, where LID and ML are, respectively, the LID and the mode lock signals. The latter signal is used as privileged information, i.e. it is known at training time but it is not available at prediction time. The only signal used at prediction time is the LID one and the prediction is accomplished according to equation (6). Again the sampling period $\tau = 2$ ms is the period of the JET real-time data network.

Figure 10 shows how the disruptive and non-disruptive regions determined by the SVM+ classifier after anomaly detection (this first supervised classifier is to be compared to the one of figure 6). The red line shows the separation frontier between the two regions in the LID space of consecutive samples. Such frontier is determined by the equation (5), which is the solution of the optimization problem defined by equation (4) with its corresponding linear restrictions. The kernels $H(\mathbf{x}_i, \mathbf{x}_j)$ and $H^*(\mathbf{x}_i^*, \mathbf{x}_j^*)$ of equation (4) for the respective LID and ML spaces are also RBF kernels. As in the precedent sections, the resulting SVM+ model has been trained to separate the training samples without error.

Table 4 shows the results of the SVM+ predictor. Again, they are obtained running the predictor as it would be in real life conditions. Figure 11 shows the evolution of the success rate and the false alarm rate with the chronological production of discharges together with the distribution of warning times.

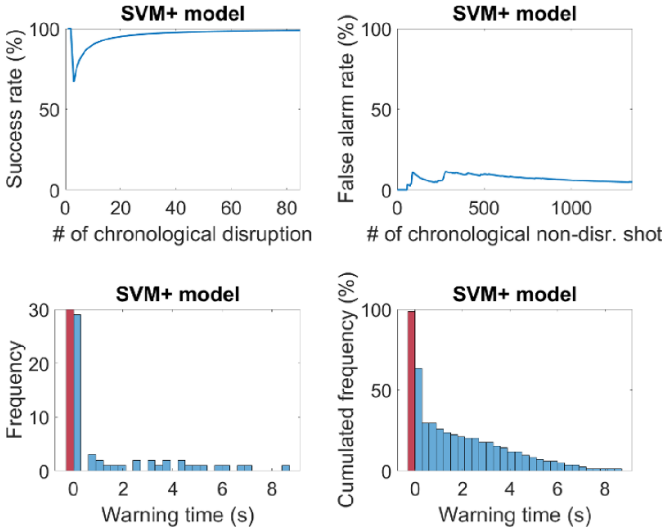


Figure 11. Top left: success rate evolution. Top right: false alarm rate evolution. Bottom left: number of correct predictions and the corresponding warning time (bin width is 300 ms). Bottom right: cumulated fraction of detected predictions at the different warning times (again, the bin width is 300 ms). The bins in red represent tardy alarms. The rate of premature alarms is about 30%.

In the case of no re-training, a direct comparison with the results of section 5.1 allows concluding that the SVM+ classifier provides slightly better results than the standard SVM model. The use of the mode lock signal as privileged information yields higher global success rate (98.81% vs 95.24%), higher success rate with positive warning time (63.10% vs 45.24%), less fraction of tardy alarms (35.71% vs 50.00%) together with less than half the false alarm rate (5.04% vs 11.53%). However, although better performances are obtained, the results are not good enough and both the tardy alarms and the false alarms have to be reduced.

With regard to the comparison of warning time distributions, it should be noted a very similar behaviour between sections 5.1 and 5.3, although a bit better for the latter. The use of privileged information to train a supervised classifier results in fewer premature alarms, which reduces the average warning times from 3.3 s to 1.7 s. However, the standard deviation remains high: 2.3 s in the present case, to be compared with 3.4 s without the use of privileged information. Concerning the distributions of tardy alarms, the delay in recognising disruptions and the standard deviations are analogous: 6 ms vs 5 ms and 4 ms vs 5 ms respectively.

A last point to emphasise is the fact that SVM+ also concentrates the alarms around the disruptions and it confirms that the LID signal is a good option to predict disruptions.

5.4. Results with privileged information and adaptive retraining of predictors

This section uses privileged information as section 5.3. The starting point is the model shown in figure 10. Again full real life and real-time situations are reproduced; the predictors are re-trained after either missed alarms or tardy alarms (this

Table 5. Predictor with privileged information and adaptive re-training. SR: success rate. SRPWT: success rate with positive warning time. TA: Tardy alarms. FA: false alarms. AWT: average warning time. STD: standard deviation of the warning times.

	Rates	AWT (s)	STD (s)
SR	98.81% (83/84)	—	—
SRPWT	90.48% (76/84)	0.112	0.481
TA	8.33% (7/84)	0.031	0.069
FA	1.15% (15/1310)	—	—

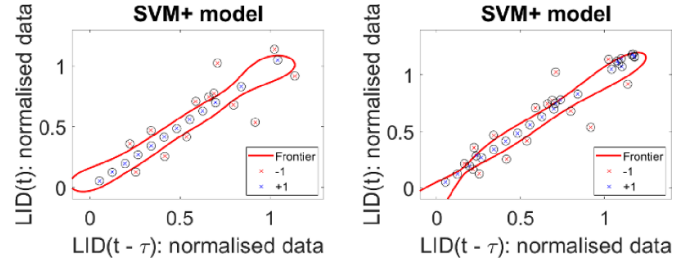


Figure 12. Resulting models after some of the re-trainings. As can be seen from figure 10 and the present plots, the effect of the adaptive re-trainings is to narrow the non-disruptive zone around the diagonal.

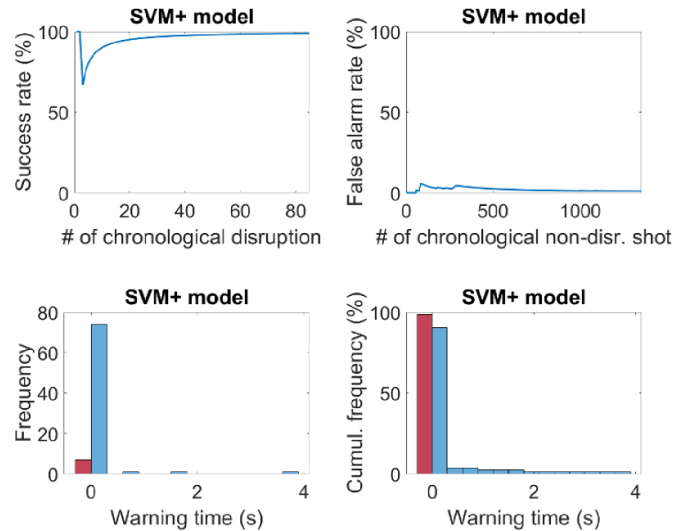


Figure 13. Top left: success rate evolution. Top right: false alarm rate evolution. Bottom left: number of correct predictions and the corresponding warning time (bin width is 300 ms). Bottom right: cumulated fraction of detected predictions at the different warning times (again, the bin width is 300 ms). The bins in red represent tardy alarms. The rate of premature alarms is about 3%.

is the difference with the previous section). The outcomes of using privileged information and adaptive re-training are summarised in table 5. There are 7 tardy alarms and 1 missed alarm and, therefore, 8 re-trainings are carried out. Examples of adaptive re-trainings are shown in figure 12.

The temporal evolution of both success rate and false alarm rate can be seen in figure 13. Moreover, the distribution of

Table 6. Summary of the four use cases. LID means the use of a LID signal. NR/R signifies (no retraining)/(use of adaptive re-trainings). NPI/PI(ML) states (no use of privileged information)/(use of the ML as privileged information). #RET represents the number of adaptive re-training as consequences of missed or tardy alarms. The rest of acronyms are the same of the previous tables.

	LID NR NPI	LID NR PI (ML)	LID R NPI	LID R PI (ML)
SR	95.24% (80/84)	98.81% (83/84)	98.81% (83/84)	98.81% (83/84)
SRPWT	45.24% (38/84)	63.10% (53/84)	88.10% (74/84)	90.48% (76/84)
TA	50.00% (42/84)	35.71% (30/84)	10.71% (9/84)	8.33% (7/84)
FA	11.53% (151/1310)	5.03% (66/1310)	2.14% (28/1310)	1.15% (15/1310)
AWT	3.3 s	1.7 s	0.6 s	0.1 s
STD	3.4 s	2.3 s	1.6 s	0.5 s
#RET			10	8

warning times and the cumulative fraction of detected disruptions are shown.

A direct comparison between non using privileged information (section 5.2) and using it (present section) shows that the success rate with positive warning time increases from 88.10% to 90.48%. Another interesting result is that the false alarm rate decreases of about a factor of 2 (from 2.14% to 1.15%).

The reduction of false alarms using privileged information means that the predictor is able to include more knowledge about what a non-disruptive behaviour is, whose more direct consequence is the reduction of premature alarms (see figures 8 and 13). The average warning time decreases from 0.558 s and a standard deviation of 1.641 s to 0.112 s and a standard deviation of 0.481 s.

A LID signal together with the use of the mode lock in the space of consecutive samples trigger alarms closer to the disruption (fewer premature alarms), rendering the predictors particularly useful for mitigation purposes.

6. Discussion

Adaptive predictors (starting from scratch), using privileged information from the mode lock, can detect disruptions in real time with high performances on the basis of only a LID signal. Four different approaches have been tested with the same dataset of 1437 JET discharges with the C wall. In all of them, the first supervised model is generated after the first disruption. The training examples of the first supervised model are made up of a very limited set that only contains 11 disruptive examples and 11 non-disruptive examples. According to table 6, the assessment of comparable predictors (columns 2–3 and columns 4–5 respectively) is always favourable to the ones that use privileged information. The adaptive predictor from scratch that takes advantage of privileged information outperforms all the others in terms of success rate with positive warning times and false alarm rates. Also the number of premature alarms has decreased, which yields average warning times $O(100 \text{ ms})$.

The effect of building adaptive predictors is to narrow the non-disruptive zone around the diagonal in the space of consecutive samples (figures 6 and 9 for standard SVM and figures 10 and 12 for SVM+). This increases the sensitivity of the predictors. But, contrary to the intuition, this augmented sensitivity does not increase the false alarm rate. Moreover,

the use of privileged information speeds up the learning process: less re-trainings are necessary for the same database of discharges. Finally, it should be emphasised that the LUPi paradigm, applied to disruption prediction in spaces of consecutive samples allows for the development of very simple classifiers just based on a couple of signals. The use of the ML signal, which is the most important one for mitigation purposes, is not necessary in real-time predictions. The knowledge provided by this signal can be used off-line at training time.

Due to the foreseen unavailability of the mode lock signal in real time in the first operation phases of JT-60SA, the combination of a LID signal together with the ML signal as privileged information constitutes a very competitive recipe for reliable disruption prediction.

An important comment about the real-time character of the predictions can be necessary. In the article, only one SVM classifier and only one signal is required. The APODIS disruption predictor uses 14 features from 7 signals and 4 SVM classifiers. The time to make predictions with APODIS is about $300 \mu\text{s}$ [15]. Therefore, in the present case, if the sampling period of the LID signal is greater than hundreds of microseconds, the real-time operation is guaranteed.

In terms of future developments, some improvements are already under study. First, anomaly detection techniques are being investigated to predict even the first disruption, so that disruptive samples are not really needed to train the first generation of SVM+ classifiers. It should also be remembered that most of the false and premature alarms are consequence of fringe jumps in the LID signal. To what extent privileged information can allow discriminating whether or not a jump in the LID signal is related to an incoming disruption is a matter under assessment. It must also be emphasised again that the proposed approach implemented by SVM+ is fully general and can be applied to any combination of signals. The deployment of the same technology to prevention and avoidance is an obvious next step in this line of research.

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation under the Project Nos. PID2019-108377RB-C31, PID2019-108377RB-C32, PID2022-137680OB-C31 and PID2022-137680OB-C32.

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

ORCID iDs

J. Vega  <https://orcid.org/0000-0002-1622-3984>
 S. Dormido-Canto  <https://orcid.org/0000-0001-7652-5338>
 R. Castro  <https://orcid.org/0000-0003-4199-9335>
 J.D. Fernández  <https://orcid.org/0009-0007-0855-914X>
 A. Murari  <https://orcid.org/0000-0002-3932-3865>

References

- [1] Boozer A.H. 2012 Theory of tokamak disruptions *Phys. Plasmas* **19** 058101
- [2] Boozer A.H. 2017 Runaway electrons and ITER *Nucl. Fusion* **57** 056018
- [3] Vega J., Murari A., Dormido-Canto S., Rattá G.A. and Gelfusa M. (JET Contributors) 2022 Disruption prediction with artificial intelligence techniques in tokamak plasmas *Nat. Phys.* **18** 741–50
- [4] Moreno R., Vega J., Dormido-Canto S., Pereira A. and Murari A. (JET Contributors) 2016 Disruption prediction on JET during the ILW experimental campaigns *Fusion Sci. Technol.* **69** 485–94
- [5] Cannas B., Fanni A., Marongiu E. and Sonato P. 2004 Disruption forecasting at JET using neural networks *Nucl. Fusion* **44** 68–76
- [6] Murari A., Vagliasindi G., Arena P., Fortuna L., Barana O. and Johnson M. (JET EFDA Contributors) 2008 Prototype of an adaptive disruption predictor for JET based on fuzzy logic and regression trees *Nucl. Fusion* **48** 035010
- [7] Vega J., Murari A., Dormido-Canto S., Hernández F., Cruz T., Gadariya D. and Rattá G.A. (JET Contributors) 2020 A linear equation based on signal increments to predict disruptive behaviours and the time to disruption on JET *Nucl. Fusion* **60** 026001
- [8] Hu W.H. et al 2021 Real-time prediction of high-density EAST disruptions using random forest *Nucl. Fusion* **61** 066034
- [9] Vega J., Murari A., Dormido-Canto S., Moreno R., Pereira A. and Acero A. (JET-EFDA Contributors) 2014 Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks *Nucl. Fusion* **54** 123001
- [10] Moreno R., Vega J. and Dormido-Canto S. (JET Contributors) 2016 *Conformal Prediction of Disruptions from Scratch: Application to an ITER Scenario (Lecture Notes in Artificial Intelligence)* vol 9653 (Springer) pp 67–74
- [11] Rea C., Granetz R.S., Montes K., Tinguely R.A., Eidietis N., Hanson J.M. and Sammulu B. 2018 Disruption prediction investigations using machine learning tools on DIII-D and Alcator C-Mod *Plasma Phys. Control. Fusion* **60** 084004
- [12] Rattá G.A., Vega J., Murari A. and Johnson M. JET-EFDA Contributors 2008 Feature extraction for improved disruption prediction analysis at JET *Rev. Sci. Instrum.* **79** 10F328
- [13] Sias G., Aleda R., Cannas B., Delogu R.S., Fanni A., Murari A. and Pau A. A multivariate analysis of disruption precursors on JET and AUG *1st IAEA Technical Meeting on Fusion Data Processing, Validation and Analysis (Nice, 1–3 June 2015)* (available at: <https://irfm.cea.fr/TMFDPVA15/>)
- [14] Esquebri S., Vega J., Murari A., Ruiz M., Barrera E., Dormido-Canto S., Felton R., Tsalas M. and Valcarcel D. (JET Contributors) 2018 Real-time implementation in JET of the SPAD disruption predictor using MARTE *IEEE Trans. Nucl. Sci.* **65** 836–42
- [15] Vega J., Dormido-Canto S., López J.M., Murari A., Ramírez J.M., Moreno R., Ruiz M., Alves D. and Felton R. (JET EFDA Contributors) 2013 Results of the JET real-time disruption predictor in the ITER-like wall campaigns *Fusion Eng. Des.* **88** 1228–31
- [16] Dormido-Canto S., Vega J., Ramírez J.M., Murari A., Moreno R., López J.M. and Pereira A. (JET EFDA Contributors) 2013 Development of an efficient real-time disruption predictor from scratch on JET and implications for ITER *Nucl. Fusion* **53** 113001
- [17] Kates-Harbeck J., Svyatkovskiy A. and Tang W. 2019 Predicting disruptive instabilities in controlled fusion plasmas through deep learning *Nature* **568** 526–31
- [18] Churchill R.M., Tobias B. and Zhu Y. 2020 Deep convolutional neural networks for multi-scale time-series classification and application to tokamak disruption prediction using raw, high temporal resolution diagnostic data *Phys. Plasmas* **27** 062510
- [19] Ferreira D.R., Carvalho P.J. and Fernandes H. 2019 Deep learning for plasma tomography and disruption prediction from bolometer data *IEEE Trans. Plasma Sci.* **48** 36–45
- [20] Guo B.H. et al 2020 Disruption prediction using a full convolutional neural network on EAST *Plasma Phys. Control. Fusion* **63** 025008
- [21] Guo B.H. et al 2021 Disruption prediction on EAST tokamak using a deep learning algorithm *Plasma Phys. Control. Fusion* **63** 115007
- [22] Zhu J.X., Rea C., Montes K., Granetz R.S., Sweeney R. and Tinguely R.A. 2020 Hybrid deep-learning architecture for general disruption prediction across multiple tokamaks *Nucl. Fusion* **61** 026007
- [23] Guo B.H. et al 2023 Disruption prediction on EAST with different wall conditions based on a multi-scale deep hybrid neural network *Nucl. Fusion* **63** 094001
- [24] Zhang Y., Pautasso G., Kardaun O., Tardini G. and Zhang X.D. 2011 Prediction of disruptions on ASDEX upgrade using discriminant analysis *Nucl. Fusion* **51** 063039
- [25] Gerhardt S.P. et al 2013 Detection of disruptions in the high- β spherical torus NSTX *Nucl. Fusion* **53** 063021
- [26] Rea C., Montes K.J., Pau A., Granetz R.S. and Sauter O. 2020 Progress toward interpretable machine learning-based disruption predictors across tokamaks *Fusion Sci. Technol.* **76** 912–24
- [27] Murari A., Lungaroni M., Peluso E., Gaudio P., Vega J., Dormido-Canto S., Baruzzo M. and Gelfusa M. (JET Contributors) 2018 Adaptive predictors based on probabilistic SVM for real time disruption mitigation on JET *Nucl. Fusion* **58** 056002
- [28] Vega J., Hernández F., Dormido-Canto S., Isayama A., Joffrin E., Matsunaga G. and Suzuki T. 2019 Assessment of linear disruption predictors using JT-60U data *Fusion Eng. Des.* **146** 1291–4
- [29] Murari A., Lungaroni M., Gelfusa M., Peluso E. and Vega J. (JET Contributors) 2019 Adaptive learning for disruption prediction in non-stationary conditions *Nucl. Fusion* **59** 086037
- [30] Murari A., Rossi R., Peluso E., Lungaroni M., Gaudio P., Gelfusa M., Rattá G. and Vega J. 2020 On the transfer of adaptive predictors between different devices for both mitigation and prevention of disruptions *Nucl. Fusion* **60** 056003

- [31] Gadariya D., Vega J., Stuart C., Rattá G.A., Card P., Murari A. and Dormido-Canto S. (JET Contributors) 2022 Performance analysis of the centroid method predictor implemented in the JET real time network *Plasma Phys. Control. Fusion* **64** 114003
- [32] Rossi R., Gelfusa M., Flanagan J. and Murari A. 2022 Development of robust indicators for the identification of electron temperature profile anomalies and application to JET *Plasma Phys. Control. Fusion* **64** 045002
- [33] Rossi R. et al 2023 A systematic investigation of radiation collapse for disruption avoidance and prevention on JET tokamak *Matter Radiat. Extremes* **8** 046903
- [34] de Vries P.C., Johnson M.F., Alper B., Buratti P., Hender T.C., Koslowski H.R. and Riccardo V. (JET EFDA Contributors) 2011 Survey of disruption causes at JET *Nucl. Fusion* **51** 053018
- [35] Snipes J.A. et al 1988 Large amplitude quasi-stationary MHD modes in JET *Nucl. Fusion* **28** 1085–97
- [36] Nave M.F.F. and Wesson J.A. 1990 Mode locking in tokamaks *Nucl. Fusion* **30** 2575–83
- [37] Rattá G.A., Vega J. and Murari A. (JET EFDA Contributors) 2014 Simulation and real-time replacement of missing plasma signals for disruption prediction: an implementation with APODIS *Plasma Phys. Control. Fusion* **56** 114004
- [38] Peluso E., Rossi R., Murari A., Gaudio P. and Gelfusa M. 2020 Alternative detection of $n = 1$ modes slowing down on ASDEX upgrade *Appl. Sci.* **10** 7891
- [39] Vapnik V. and Vashist A. 2009 A new learning paradigm: learning using privileged information *Neural Netw.* **22** 544–57
- [40] Kamada Y., Di Pietro E., Hanada M., Barabaschi P., Ide S., Davis S., Yoshida M., Giruzzi G. and Sozzi C. (the JT-60SA Integrated Project Team) 2022 Completion of JT-60SA construction and contribution to ITER *Nucl. Fusion* **62** 042002
- [41] Craciunescu T. and Murari A. 2022 Detection of changes in the dynamics of thermonuclear plasmas to improve the prediction of disruptions *Nonlinear Dyn.* **111** 1–15
- [42] Chen J. et al 2014 High resolution polarimeter-interferometer system for fast equilibrium dynamics and MHD instability studies on Joint-TEXT tokamak *Rev. Sci. Instrum.* **85** 11D303
- [43] Murari A., Zabeo L., Boboc A., Mazon D. and Riva M. 2006 Real-time recovery of the electron density from interferometric measurements affected by fringe jumps *Rev. Sci. Instrum.* **77** 073505
- [44] Cherkassky V. and Mulier F. 2007 *Learning from Data. Concepts, Theory, and Methods* 2nd edn (Wiley)