

## PAPER

# Real-time prediction of high-density EAST disruptions using random forest








To cite this article: W.H. Hu *et al* 2021 *Nucl. Fusion* **61** 066034

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

## You may also like

- [Small target detection and window adaptive tracking based on continuous frame images in visible light background](#)  
Zheng Wang, Jianping Zeng, Xiaoli Xie et al.
- [Individual autonomous safety intelligence technology for radioactive material transportation—Multi-sensor fusion early warning technology based on evidence theory](#)  
Qin Yang, Zhenhai Liu, Xiaohua Yang et al.
- [Free-response operator characteristic models for visual search](#)  
T P Hutchinson

# Real-time prediction of high-density EAST disruptions using random forest

W.H. Hu<sup>1</sup>, C. Rea<sup>2,\*</sup>, Q.P. Yuan<sup>1</sup>, K.G. Erickson<sup>3</sup>, D.L. Chen<sup>1</sup>, B. Shen<sup>1</sup>, Y. Huang<sup>1</sup>, J.Y. Xiao<sup>4</sup>, J.J. Chen<sup>4</sup>, Y.M. Duan<sup>1</sup>, Y. Zhang<sup>1</sup>, H.D. Zhuang<sup>1</sup>, J.C. Xu<sup>5</sup>, K.J. Montes<sup>2</sup>, R.S. Granetz<sup>2</sup>, L. Zeng<sup>1</sup>, J.P. Qian<sup>1</sup>, B.J. Xiao<sup>1</sup>, J.G. Li<sup>1</sup> and EAST Team

<sup>1</sup> Institute of Plasma Physics, Chinese Academy of Science, Hefei, China

<sup>2</sup> Massachusetts Institute of Technology, Plasma Science and Fusion Center, Cambridge, MA United States of America

<sup>3</sup> Princeton Plasma Physics Laboratory, Princeton, NJ United States of America

<sup>4</sup> School of Nuclear Science and Technology, University of Science and Technology of China, Hefei, China

<sup>5</sup> School of Mechanical Engineering, Anhui University of Science and Technology, Huainan, China

E-mail: [crea@mit.edu](mailto:crea@mit.edu)

Received 26 January 2021, revised 6 April 2021

Accepted for publication 13 April 2021

Published 18 May 2021



## Abstract

A real-time disruption predictor using random forest was developed for high-density disruptions and used in the plasma control system (PCS) of the EAST tokamak for the first time. The disruption predictor via random forest (DPRF) ran in piggyback mode and was actively exploited in dedicated experiments during the 2019–2020 experimental campaign to test its real-time predictive capabilities in oncoming high-density disruptions. During dedicated experiments, the mitigation system was triggered by a preset alarm provided by DPRF and neon gas was injected into the plasma to successfully mitigate disruption damage. DPRF's average computing time of  $\sim 250 \mu\text{s}$  is also an extremely relevant result, considering that the algorithm provides not only the probability of an impending disruption, i.e. the disruptivity, but also the so-called feature contributions, i.e. explainability estimates to interpret in real time the drivers of the disruptivity. DPRF was trained with a dataset of disruptions in which the electron density reached at least 80% of the Greenwald density limit, using the zero-dimensional signal routinely available to the EAST PCS. Through offline analysis, an optimal warning threshold on the DPRF disruptivity signal was found, which allows for a successful alarm rate of 92% and a false alarm rate of 9.9%. By analyzing the false alarm causes, we find that a fraction ( $\sim 15\%$ ) of the misclassifications are due to sudden transitions of plasma confinement from H- to L-mode, which often occur during high-density discharges in EAST. By analyzing DPRF feature contributions, it emerges that the loop voltage signal is that main cause of such false alarms: plasma signals more apt to characterize the confinement back-transition should be included to avoid false alarms.

Keywords: disruption prediction, EAST, real time, mitigation

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

## 1. Introduction

One of the most critical parameters that affect high-fusion performance in tokamak reactors is the plasma density.

\* Author to whom any correspondence should be addressed.

Empirical scalings from experimental observations across various devices show that when the plasma density increases up to near the Greenwald density ( $n_{\text{GW}}$ ) limit, a stability threshold is reached and the plasma will transition to unstable operational regimes or poor confinement modes, with the

final consequence of a disruption [1–3]. According to ITER’s design parameters, the international experimental reactor will operate at  $n_e/n_{GW} = 0.85$  [4], where  $n_e$  is the plasma electron density, thus making ITER’s operational scenario susceptible to density-limit disruptions. Plasma disruptions are associated with the sudden degradation of particle confinement and loss of stored energy, which can eventually cause massive heat and electromagnetic loads to be dumped onto the device’s wall components on time scales ranging from micro- to milliseconds. Certain clear signatures of density-limit disruptions, such as magnetohydrodynamics (MHD) instabilities as  $n = 1$  modes, can be observed with enough alarm time to only mitigate the disruption consequences and not just avoid this dangerous chain of events. Therefore, investigating possible alternative ways to predict density-limit disruptions on existing tokamak devices becomes valuable.

For example, path-oriented approaches have recently been explored in ASDEX Upgrade and TCV tokamaks to design trajectory optimization strategies aimed at sustaining plasma operations as long as possible at full performance, thus avoiding disruptions [5]. In addition to other physics-based predictive models [6], machine learning methodologies have boosted the efforts to investigate and predict density-limit disruptions: on the ADITYA tokamak, a multi-layer perceptron, i.e. a fully connected feed-forward neural network model, was originally used in the early 2000s to train a density-limit disruption predictor. The authors collected 2000 samples from 23 discharges to train the predictor and an alarm time of 0.35–56 ms was achieved on the test discharges with an alarm threshold of 0.94 [7]. More recently, on the J-TEXT tokamak, a neural network-based model was also used to train a density limit predictor, and a successful alarm rate of 82.8% with a false alarm rate of 12.3% was obtained [8]. To further improve these performances, the authors subsequently devised a two-stage hybrid neural network with an increased successful alarm rate higher than 90% and with a false alarm rate lower than 10%. Eventually, such model was implemented in the density control system to be used for real-time density-limit disruption alarm ( $T_{\text{alarm}} \sim 30$  ms) and mitigation [9].

The above examples are among many showing the great potential of data-driven models in disruption prediction research: references [3–28] in [10] present a recently updated literature review, although disruption prediction research is constantly being improved upon thanks to newly published results. This is also correlated with the availability of more efficient computational resources and the exploitation of larger sets of training data, which allows for further optimization of data-driven algorithm results, i.e. a higher successful prediction rate and lower false prediction rate.

Random forest models have already been used to investigate possible disruption prediction solutions in DIII-D and Alcator C-Mod [11], reaching an overall accuracy greater than  $\sim 97\%$ . This analysis was subsequently extended also to EAST [12], where differences in performances have been more closely analyzed using training databases of major disruptions that were not discriminated by cause. A comparative analysis of the real-time implementation of the disruption prediction via random forest algorithm (DPRF) can be found in Rea

*et al* [13]. As random forest is a supervised learning method, data samples need to be labeled according to physics knowledge before training. The density-limit boundary in tokamak plasmas is related to multiple parameters and various physics processes, including plasma energy loss, particle confinement degradation, plasma cooling and the rise of impurity radiation at the edge [3]. To properly identify density-limit disruptions, one should investigate the disruption paths in all collected high-density disruptions in EAST to find dynamics similar to the increasing plasma density, plus the rise in impurity radiation from the edge, plus edge cooling and steepening of the current profile in the vicinity of the  $q = 2$  surface, together with destabilization of  $n = 1$  MHD modes (principally those with poloidal mode number  $m = 2$ ), which eventually grow and produce a major disruption. This is a rather complex effort, quite difficult to physically diagnose at scale. In this paper we consider high-density disruptions that include density-limit cases, but might also contain spurious dynamics such as uncontrolled density ramps, density feedback failures, etc, and we present a real-time data-driven predictive model for disruption warning in EAST (DPRF) based on the random forest algorithm [14].

DPRF integration in the EAST plasma control system (PCS) is the first step in augmenting emergency response alarms to oncoming disruptions. But the final aim of disruption research should be to continuously provide information on the current plasma state to monitor its proximity to stability limits [15], and therefore provide avoidance strategies to extend safe plasma operations. In order to do so, interpretability results are a key feature of machine learning algorithms: by interpreting the drivers of the increasing probability of an impending disruption in real time, the PCS can be informed on which actuators to use to steer the plasma away from the instabilities. DPRF indeed provides explainable predictions through the feature contribution analysis, which will be introduced in section 2 and used to better analyze and interpret DPRF performances.

The manuscript is organized as follows: section 2 introduces the dataset used to train the predictor, with several details about the algorithm, the offline optimization process, the database collection and sub-selection on the basis of the Greenwald density limit. Section 3 describes the implementation of the predictor into the real-time PCS in EAST and reports on specific results of dedicated closed-loop experiments where the predictor is used to trigger the massive gas injection (MGI) system to mitigate the disruption. An example of false alarm dynamics is also described in this section. In section 4 we discuss the overall performances of the real-time predictor for discharges where DPRF ran at different thresholds on the disruptivity. Finally, we summarize the main results and provide an outlook for future work in section 5.

## 2. Real-time disruption prediction model

### 2.1. Database collection

In order to develop a data-driven algorithm, it is essential to gather enough data to train the machine learning algorithm

of choice. We focused on EAST's experimental campaigns between 2015 and 2018, from which 483 disruptive discharges were selected based on their line-averaged density value ramping up to or above the 80% Greenwald density limit during the flat-top phase of the plasma current. As a reminder, the Greenwald density limit is calculated as

$$n_{\text{GW}}[10^{20} \text{ m}^{-3}] = \frac{I_p [\text{MA}]}{\pi a^2 [\text{m}^2]}, \quad (1)$$

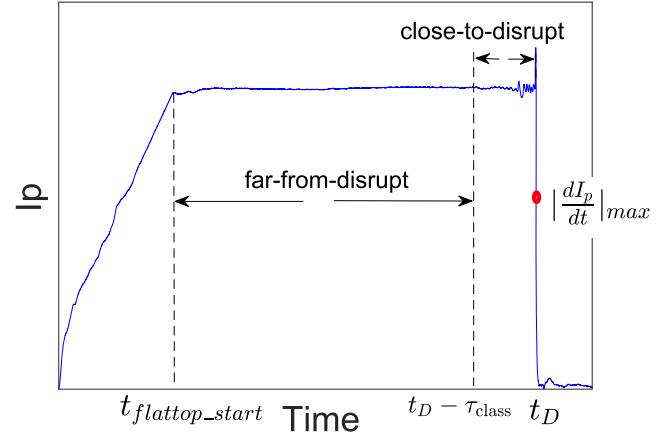
where  $I_p$  is the plasma current and  $a$  is the minor radius. The plasma line-averaged density comes from the diagnostic of the hydrogen cyanide (HCN) interferometer [16]. To complement the disruptive subset, 483 non-disruptive discharges are randomly selected from a variety of scenarios during the same experimental campaigns, and together with the disruptive shots they compose a training set of 966 EAST discharges. For the application reported in this manuscript, we rely on the SQL database developed for disruption prediction studies in EAST [12], which gathers numerous plasma signals sampled at 10 Hz during the flat-top phase, while additional sampling at 100 Hz occurs for the last 250 ms period before each disruption. This non-uniform sampling rate is motivated by the univariate analysis on several EAST plasma signals, and aimed at capturing high-frequency information relevant to the oncoming disruptions, although for several diagnostic signals this implies a downsampling. Since DPRF is not a sequence-based algorithm, all training samples are evaluated independently from their correlations in time, plus no time derivatives are included among the input features. All training samples are therefore taken from the flat-top phase of EAST discharges: a typical flat-top duration is around 8–10 s, though some discharge has reached 100 s [17].

As the chosen data-driven algorithm is a supervised one, we also need to explicitly assign labels to the samples in our dataset. For disruptive shots, as shown in figure 1 and in reference to [11], the samples between the beginning of the flat-top phase and  $t_D - \tau_{\text{class}}$  are regarded as 'stable', while samples between  $t_D - \tau_{\text{class}}$  and  $t_D$  are labeled as 'unstable'. Here,  $t_D$  is the disruption time and it is defined to be the time of  $|dI_p/dt|_{\text{max}}$ .  $\tau_{\text{class}}$  is therefore defined with respect to the time of the disruption,  $t_D$ .

The choice of  $\tau_{\text{class}}$  directly influences the performances of DPRF: the  $\tau_{\text{class}}$  optimal value was chosen on the basis of a  $K$ -fold cross-validation technique as documented in [12], in order to minimize the false alarms and the missed warnings on a shot-by-shot basis on the validation set. Also in this case, a cost-sensitive binary classification metric called the  $F_\gamma$ -score is adopted to evaluate DPRF's performance on the validation set, which is calculated as

$$F_\gamma = (1 + \gamma^2) \frac{\text{precision} \cdot \text{recall}}{\gamma^2 \cdot \text{precision} + \text{recall}}. \quad (2)$$

Here,  $\text{precision} = \text{TP}/(\text{TP} + \text{FP})$  and  $\text{recall} = \text{TP}/(\text{TP} + \text{FN})$ , where TP (true positive) means disruptive shots that are correctly classified as disruptive, while FP (false positive) are non-disruptive shots that are misclassified as disruptive. TN (true negative) are non-disruptive shots that are correctly

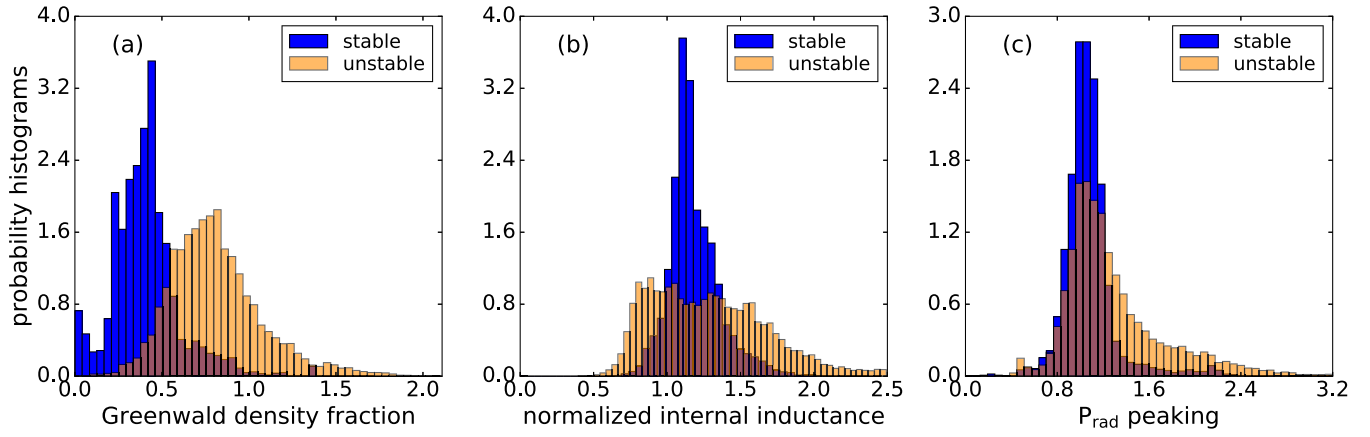


**Figure 1.** The scheme of stable ('far-from-disrupt') and unstable ('close-to-disrupt') samples for a typical disruptive discharge.  $t_D$  is the time of disruption, which is defined as the time of  $|dI_p/dt|_{\text{max}}$ .

classified as non-disruptive, and FN (false negative) are disruptive shots that are misclassified as non-disruptive.  $\gamma$  is chosen according to the operational needs; for example,  $F_\gamma = \text{precision}$  when  $\gamma = 0$  and  $F_\gamma = \text{recall}$  when  $\gamma \rightarrow \infty$ . To compare  $F_\gamma$  for a set of  $\tau_{\text{class}}$  values the following steps are taken: (i) dividing all training samples into  $K$  subsets, of which  $K - 1$  subsets act as training samples and the remaining subset acts as testing samples, or a validation set; (ii) with every subset taking turns to be the testing set, the random forest is trained repeatedly  $K$  times for each value of  $\tau_{\text{class}}$ . Both the  $F_1$  and  $F_2$  score are checked with  $\gamma = 1$  or 2. The  $F_2$  score has a higher accuracy on the disruptive class (the minority class), assigning a higher cost to those misclassifications. The  $\tau_{\text{class}}$  was found not to differ too much between the  $F_1$  and  $F_2$  scores, and the final optimal value of  $\tau_{\text{class}}$  was found to be 1.7 s with respect to the disruption time  $t_D$ .

For the non-disruptive shots, all of the flat-top samples are collected and labeled as 'stable'. During the training process, 'unstable' and 'stable' labels are encoded as numbers, either 0 or 1. The binary classification choice at the core of the DPRF algorithm is supported also by the univariate analysis of EAST plasma signals relevant to the disruption dynamics we are interested in predicting. In particular, figure 2 shows the probability histograms for the Greenwald density fraction, the normalized internal induction and the peaking factor for the radiation profile for stable and unstable samples in our dataset. Here, we define the peaking factor as the ratio of the average of the  $P_{\text{rad}}$  signals of the AXUV core channels to the average of all the channels, excluding those that look in the divertor region. The core channels are fixed as the six centralmost chords out of the available 64. An updated definition of the  $P_{\text{rad}}$  peaking that follows more closely what the authors have done in DIII-D [18] is already available offline and should be retroactively applied to the EAST experimental database for disruption prediction studies. This is part of our next step, aimed at upgrading DPRF on EAST.

In figure 2(a), around half of the unstable samples are of  $\text{GW}_{\text{frac}} > 0.8$ , while most stable samples are of  $\text{GW}_{\text{frac}} < 0.5$ , which is in agreement with the data selection



**Figure 2.** The probability histograms of the (a) Greenwald density fraction ( $GW_{frac}$ ), (b) normalized internal inductance and (c) peaking factor of the radiation profile are shown for the training set, including time slices of both ‘stable’ and ‘unstable’ samples. The vertical axis is normalized so that the integral under each histogram is unity.  $P_{rad}$  peaking is not included in DPRF’s input features due to the unavailability of real-time digitization of radiation measurement, but its offline analysis supports edge-cooling mechanisms associated with density-limit disruptions.

method mentioned above. The normalized internal inductance signal contains information on the current density profile distribution: the histograms in figure 2(b) show a different behavior for stable and unstable samples. The latter distribution is rather less peaked than the stable one, with a tendency to a bimodal behavior around  $li \sim 0.7$  and  $li \sim 1.3$ . The tail of unstable samples with  $li > 1.5$  implies a peaked distributed current density, but the low- $li$  branch hints at a spurious training dataset in terms of disruption dynamics. Figure 2(c) shows the radiation peaking factor for stable and unstable samples: the histogram shows that unstable samples have a higher peaking factor value, which is consistent with the density-limit behavior of plasma edge cooling. However, the conflicting behavior of the normalized internal inductance distribution shows that more evidence would be needed to fully consider the collected high-density disruption discharges in EAST as density-limit disruptions.

The signals collected for these samples are listed, together with their description, in table 1. A total of 57475 training samples is collected,  $\sim 16281$  of which are labeled as unstable. Although the training set shows a relatively strong imbalance in the class composition, no improvement was found in DPRF performances when using over-/undersampling techniques to rebalance the class composition [19]. The adoption of the  $F_\gamma$  score on a shot-by-shot basis was found to be sufficient to obtain robust DPRF performances. The signals in table 1 are all scalar (or zero-dimensional) plasma signals and all of them are available in the PCS for real-time usage.

## 2.2. Off-line training and PCS integration

DPRF is based on the random forest *shallow* machine learning algorithm [14]: the final class membership probability is obtained by developing a large number (typically hundreds) of independent, de-correlated base learners, i.e. decision tree models, thus collecting a parallel set of predictions. The final prediction from the ensemble is aggregated by averaging this very large number of model predictions, and thus reducing the

model’s bias and variance. More methodological details about random forests can be also found in [11,20].

As mentioned in section 2.1, several critical model parameters ( $\tau_{class}$ , disruptivity threshold, etc) are found offline through a nested  $K$ -fold cross-validation approach to maximize the  $F_\gamma$ -score on a shot-by-shot basis, and through a parameter grid search [12]. In particular, the DPRF model presented for this application relies on a random forest composed of 500 decision trees (individual estimators) fully grown—the architectural details (number of trees and layers per each tree) are chosen by minimizing of the out-of-bag error rate for a specific set of model parameters, as discussed in [11].

The training procedure is conducted similarly to what is detailed in references [11,12,21]: the random forest classifier core algorithm is taken *off-the-shelf* from the scikit-learn open source library<sup>6</sup> and used to train a number of individual base learners, i.e. decision trees, to classify unstable/stable samples organized by shot, and eventually the average result of these classifiers is taken as the final probability of a disruption, i.e. the disruptivity. At the same time, the *treeinterpreter* package<sup>7</sup> is used to decompose the drivers of the disruptivity following a linear regression approach into terms called bias and feature contributions. The decomposition formula for the predicted disruptivity is shown in equation (3), where  $contrib_m$  is the feature contribution from the  $m$ -th input variable and the bias comes from the sample mean of the training population and is constant.

$$\text{disruptivity} = \frac{1}{M} \sum_{m=1}^M \text{bias}_m + \sum_{m=1}^M \left( \frac{1}{M} \sum_{n=1}^N \text{contrib}_m(n) \right), \quad (3)$$

where  $M$  is the total number of trees in the forest and the  $n$  index runs on the number of the input features. In the equation above, the bias term is strongly dependent on the training set and represents the sample distribution in the class of interest.

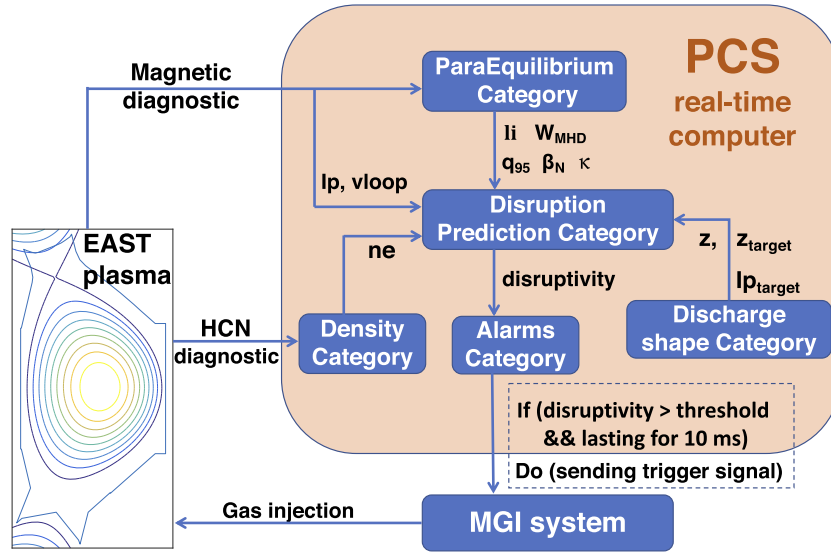
<sup>6</sup> <https://scikit-learn.org/stable/>

<sup>7</sup> <https://github.com/andosa/treeinterpreter>



**Table 1.** DPRF real-time input signals and their description.

Signal	Description
$I_{p\text{error}}$	$(I_p - I_{p\text{target}})/I_{p\text{target}}$
$V_{\text{loop}}$	Plasma loop voltage (V)
$GW_{\text{frac}}$	Greenwald fraction: ratio of plasma density to Greenwald density limit
$\beta_N$	Normalized ratio of plasma pressure to magnetic pressure
$W_{\text{MHD}}$	Plasma stored energy (J)
$l_i$	Plasma internal inductance
$\kappa$	Plasma elongation
$q_{95}$	Safety factor at 95% flux surface
$z_{\text{error}}$	$(z - z_{\text{target}})/a$ , where $z$ is the vertical position of plasma center, $a$ is the minor radius of tokamak

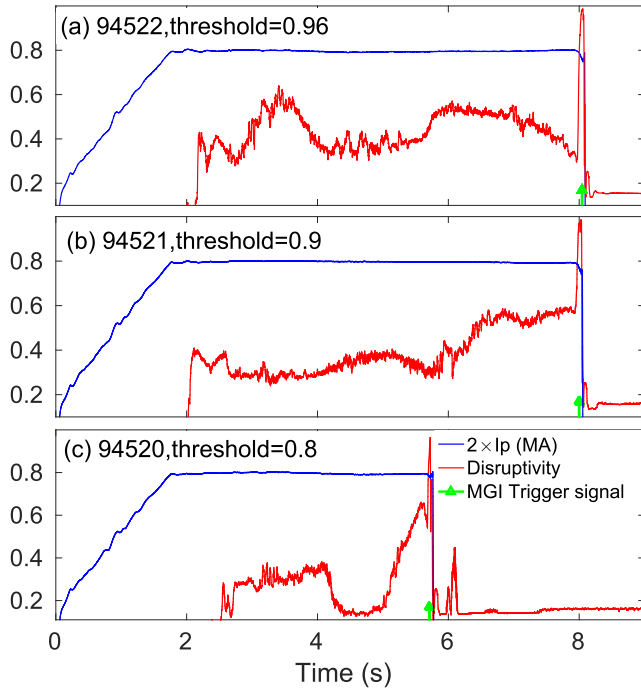


**Figure 3.** The disruption prediction category implemented in the real-time computer of EAST PCS. The input signals from different categories and magnetic diagnostic are transferred to it and its calculated disruptivity is transferred to the alarms category. The disruption prediction, alarm and MGI system make up the close loop of disruption prediction and disruption mitigation system in EAST.

Reporting the example discussed in section 4 of [20]: if the population of the samples in the available training set is distributed as 90% belonging to the non-disruptive class and 10% to the disruptive class, then the bias when evaluating the contributions for the disruptive class predictions will be 0.1, and 0.9 for the non-disruptive class predictions. The feature contributions are then calculated from the relative differences in the class-of-interest (‘disruptive’) population, after evaluating each feature in all decision paths at the inference time. Contributions are obtained from pre-computed values during the training procedure: an in-depth illustrative example is reported in section 4 of [20], beyond what is also detailed in [22].

After training, the random forest model is translated from Python into C language and then compiled as an external library with the PCS so that it can run during plasma experiments when enabled in the PCS. Based on that, a disruption prediction category is built. Here, in the PCS framework, a ‘category’ corresponds to a control function with an actuator or class of actuators, such as shape control and current control. For the disruption prediction category, part of its input signals come from direct diagnostic measurements and the other input signals are from calculated parameters of other control

categories. As shown in figure 3,  $I_p$  and  $V_{\text{loop}}$  are acquired from the magnetic diagnostics, and the plasma density  $n_e$  is acquired from the density category, where density is calculated with the HCN diagnostic [23]. The plasma equilibrium and confinement parameters including  $l_i$ ,  $W_{\text{MHD}}$ ,  $q_{95}$ ,  $\beta_N$ ,  $\kappa$  are provided by the Paraequilibrium (PEFIT) category. The actual and target plasma vertical position  $z$ ,  $z_{\text{target}}$  and target plasma current  $I_{p\text{target}}$  are from the current and shape control category [24]. The calculation time of different categories over one cycle is different and the PEFIT category has the slowest calculation time. In EAST, the real-time equilibrium is reconstructed via parallelized GPU calculations [25,26]. It takes 500  $\mu\text{s}$  for PEFIT to complete one equilibrium iteration, parameter calculation and data transfer. In addition, DPRF calculates the disruptivity with an average computing time of  $\sim 200 \mu\text{s}$ —an example is shown in the middle panel of figure 5. Therefore, the time cycle of the disruption prediction category is set to be 1 ms, which is spent on input signals reading, disruptivity value calculation and disruptivity value transferring to the alarms category. If the disruptivity goes above a preset threshold (a parameter that can be configured during experiments) the alarms category starts to record it, and if it lasts for



**Figure 4.** Plasma current signal scaled by a factor 2 (blue), disruptivity signal (red) and MGI trigger signal (green) by real-time disruption warning with different alarm thresholds (a) 0.96, (b) 0.9, (c) 0.8.

more than 10 ms, a signal is sent out to the MGI system to trigger gas injection into the plasma to mitigate the disruption effects.

### 3. Experiments using DPRF for real-time disruption prediction

#### 3.1. Cases where DPRF triggered the MGI

During the 2019–2020 EAST experimental campaign, DPRF was tested while running using different warning thresholds for the disruptivity value, beyond the optimal 0.8 coming from offline performance optimization. As shown in figure 4, three threshold values (0.8, 0.9, 0.96) are tested in real time in three different discharges to trigger the MGI system.

When DPRF runs in real-time experiments, it is usually enabled at the start of the flat-top phase because it is trained with all data taken from flat-top phases. The examples in figure 4 show that DPRF is able to predict an oncoming disruption with an alarm time that ranges around 40–60 ms, depending on the situation. The disruptivity signal is sent to the alarm category in the PCS, which triggers the MGI valve. For these experiments, the gas valve was set to receive the trigger but to not inject any gas.

The MGI trigger signal and the actual disruption time of the three shots in figure 4 are reported in table 2: it shows that when the disruptivity threshold increases from 0.8 to 0.96,  $T_{\text{alarm}}$  decreases from 57 ms to 44 ms. Intuitively, a higher disruptivity threshold results in a shorter alarm time. This further validates the offline selection of 0.8 for the disruptivity threshold

**Table 2.** Time of MGI trigger signal and disruption time of shots with real-time disruption predictor. The alarm time is  $T_{\text{alarm}} = t_D - t_{\text{trigger}}$ .  $T_{\text{alarm}}$  shortens if the disruptivity threshold is increased, thus further validating the offline selection of 0.8 for the disruptivity threshold as the optimal choice for maximizing the DPRF performance and alarm time.

Shot number	Threshold	$t_{\text{trigger}}$ (s)	$t_D$ (s)	$T_{\text{alarm}}$ (ms)
94520	0.80	5.708	5.765	57
94521	0.90	7.995	8.048	53
94522	0.96	8.040	8.084	44

as the optimal choice for maximizing the DPRF performance and the alarm time.

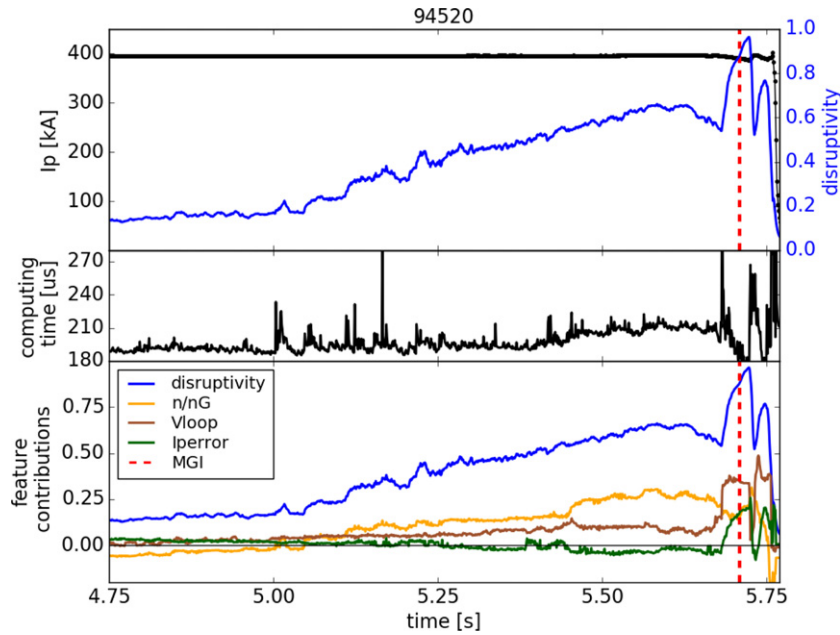
Figure 5 shows EAST shot 94520 during the last second before the disruption event. In the middle panel of figure 5(b) the DPRF computing time for each PCS cycle is shown to be around 200  $\mu\text{s}$ , which is much shorter than the actual cycle time, which is set to 1 ms. Figure 5(c) shows the three highest feature contributions to DPRF disruptivity: the Greenwald density fraction, the loop voltage and the error on the pre-programmed plasma current. At around  $t = 5.7$  s the  $V_{\text{loop}}$  contribution has a big increase and the  $I_{p_{\text{error}}}$  contribution also increases, signaling an increasing plasma resistivity and as a result, an increase in the disruptivity value to the preset threshold of 0.8. This behavior, combined with the decreased contribution from the Greenwald density fraction, hints at a disruption due to a radiative event or an impurity accumulation and not to a density-limit case. The increase in the disruptivity above the threshold for 10 ms triggers the MGI system valve in any case. Note that 10 ms is also chosen as a model parameter, together with  $\tau_{\text{class}}$  and according to reference [12].

#### 3.2. Analysis of mitigation effects when MGI is triggered by DPRF

Two reference discharges, 94537 and 94538, are used to assess the viability of using DPRF to predict disruptions and trigger the MGI to mitigate any deleterious effect. Referring to figure 6, both discharges present the electron density ramping up, with an increasing disruptivity that indicates increasing risk of disruption.

Two different disruptivity thresholds were chosen for the two discharges, as indicated in figure 6: while discharge 94537 was used to test that the DPRF alarm was properly sent to trigger the MGI, no gas was injected into the plasma, thus only mimicking the actual mitigation scheme. In the subsequent shot, discharge 94538, the threshold was lowered to aim for an earlier warning: the alarm was sent to the MGI system and triggered at around  $t = 4.09$  s, injecting 4500 Pa.L of neon gas into the plasma.

We then analyze the mitigation effects, comparing the two cases from three different aspects: (1) We compare the two discharges during thermal quench, finding that the peak of total radiation power in the bulk plasma is much higher in 94538 than 94537, as shown in figure 6(d)—consistent with the massive injection of impurities into the plasma. (2) After the disruption in figure 7(a), we show that the ion saturation current



**Figure 5.** Figure adapted from [13]: last 1 s of EAST discharge 94520. The top panel shows the disruptivity in blue, and the plasma current time traces in black. The MGI trigger signal occurs at 5.71 s (the red dashed line). The middle panel shows the computing time of each PCS cycle for the DPRF algorithm and the bottom panel shows the disruptivity together with the highest three feature contributions, which are the Greenwald fraction ( $GW_{frac}$ ,  $n/nG$  in the figure),  $V_{loop}$  and  $I_{perror}$ .

density is successfully reduced for the gas-injected case compared with the case without injection in figure 7(c). The ion saturation current density is measured by the lower outer divertor Langmuir probes. According to the vertical position of the plasma current centroid, both 94537 and 94538 disrupt near the vessel's midplane. (3) Finally, we show that the eddy currents at the lower divertor support [27] are greatly reduced in 94538 compared with 94537—see figure 7(b) versus figure 7(d).

### 3.3. Cases where DPRF false alarms trigger the MGI

DPRF is not a perfect classifier and false alarms were also recorded: discharge 94535 is an example where the MGI was triggered by a false alarm. In figure 8, we can see that the density continuously ramps up in panel (b), while in panel (a) the disruptivity simultaneously increases and when it reaches the preset threshold of 0.96 for 10 ms an alarm is sent to the MGI system at  $t_{trigger} = 4.65$  s, though without gas injection. Nevertheless, the plasma does not disrupt over the whole flat-top phase (the gray shadowed region). Even though a disruption eventually occurs during the ramp-down phase, this is still regarded as a false alarm because DPRF's region of validity for warnings coincides with the current flat-top phase, as it is trained with samples only from the flat top.

In this particular discharge (94535), the plasma transitions to the H-mode near  $t = 3$  s, when the  $D_\alpha$  signal shows that small edge-localized modes (ELMs) appear [28] (green curve in figure 8(b)). With the density increasing, the plasma stored energy decreases until around  $t = 4.65$  s, when a sudden decrease of  $W_{MHD}$  happens, indicating an H–L back-transition. In EAST discharges, H–L back-transitions are empirically found to often occur during density ramping experiments. With the density continuously ramping up, the threshold power of the H mode increases as well, and more energy is

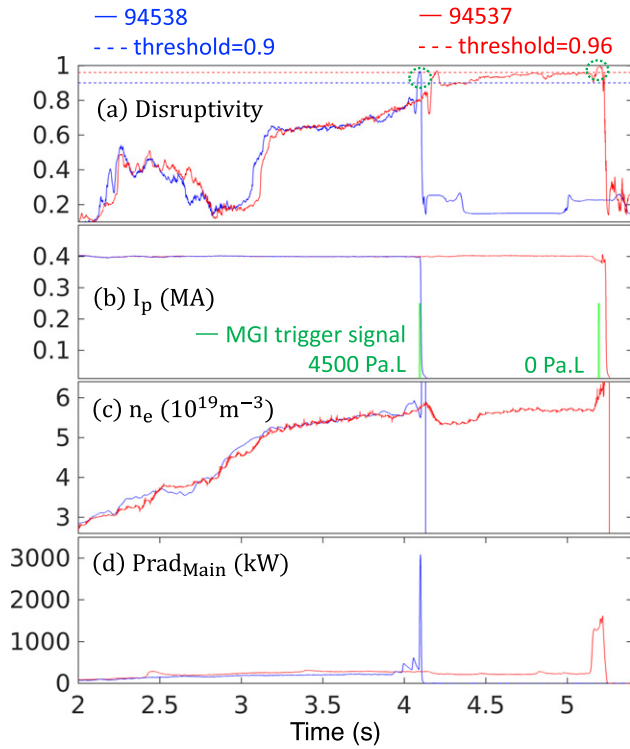
needed to sustain the H mode [29,30]. However, as shown in figure 8(c), both the lower hybrid wave (LHW) and the electron cyclotron (ECRH) heating power remain almost constant while  $W_{MHD}$  decreases constantly. At around  $t = 4.7$  s, the  $D_\alpha$  signal shows a peak and then the small ELMs disappear, implying that an H–L back-transition happens.

Still in figure 8, the bottom panel (d) shows the top three main contributing features to the disruptivity between 4.25 s and 5 s, which are the loop voltage, the Greenwald density fraction and the error on the plasma current centroid. This zoomed-in view of the feature contribution behavior shows that when approaching the back-transition moment,  $V_{loop}$  contributions increase suddenly, with a drop in the  $GW_{frac}$  and  $z_{error}$  contributions. The great deviation in the increasing  $V_{loop}$  signal causes the disruptivity to spike and the DPRF false alarm. The loop voltage, usually signaling a change in the plasma resistivity, can be influenced not only by direct changes in the bulk plasma radiation and density but also by malfunctions of the plasma heating/current driven system (hardware problem). Beside that, the  $V_{loop}$  signal might also be influenced by plasma radial motion due to a plasma's dynamical response when  $\beta_p$  suddenly changes during H–L transitions. All these causes should be taken into account when running a data-driven predictor, thus also ensuring its robustness against hardware failures. In EAST, we have empirically found that a variation of the  $V_{loop}$  signal may sometimes not represent an incumbent disruption but rather a cause of false alarms.

## 4. Overall DPRF performance analysis during 2019–2020 EAST experimental campaign

During the EAST experimental campaign that started from the Autumn of 2019 and ended in the Spring of 2020, DPRF was





**Figure 6.** (a) The disruptivity and alarm threshold (dotted line), (b) current and MGI trigger signal, (c) density and (d) total radiated power in bulk plasma evolution of two reference shots 94537 (blue line) and 94538 (red line). For shot 94538, the MGI system is triggered and neon gas with an amount of 4500 Pa.L is injected into plasma after the disruptivity value reaches the threshold value of 0.9 and lasts for 10 ms. The neon gas is injected at around  $t = 4.09$  s and immediately after that the plasma disrupts at  $t = 4.10$  s. For shot 94537, whose threshold is 0.96, the trigger signal is sent out at around  $t = 5.23$  s. In this case zero gas is injected and the plasma disrupts at  $t = 5.24$  s.

enabled in the PCS and kept running in the background during 1040 plasma discharges. In every one of these discharges DPRF was set to start working from  $t = 2$  s, to be enabled only during the flat-top phase. This allowed us to gather enough statistics to assess the overall performance of DPRF, as shown in table 3. Post-mortem analysis revealed that from a total of 1040 plasma discharges, 456 shots did not disrupt and could be categorized as healthy plasmas, therefore presenting a complete ramp-up, flat-top and ramp-down phase. In addition, according to the density signal, only 50 disruptive shots could be identified as high-density disruptions, with their Greenwald density fraction reaching at least 80% during the flat-top phase of the plasma current. Discharges that matched the following conditions were excluded from the performance evaluation: (i) shots with duration shorter than 2 s (DPRF is only enabled at  $t = 2$  s); (ii) shots with missing/malfunctioning density signal; (iii) shots disrupting without reaching 80% Greenwald density limit; (iv) shots disrupting after the end of the flat-top phase.

The metrics that we defined to characterize DPRF performances are reported in the table. A successful alarm (SA) identifies a disruptive shot with a trigger signal sent out before the end of the flat-top phase. A false alarm (FA) identifies those cases when DPRF triggered a warning before the flat-top end in

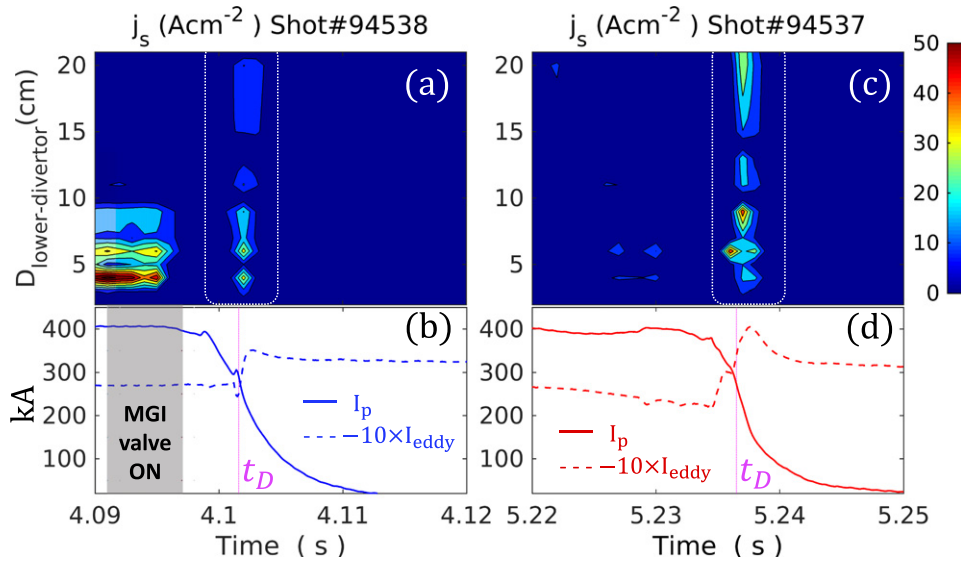
non-disruptive discharges. The SA and FA rate are normalized respectively by the number of disruptive and all non-disruptive shots.

By scanning the DPRF warning threshold from 0.7 to 0.95 while keeping the warning duration time fixed as 10 ms, we can see decreasing performances and mean alarm times with a more stringent request for a higher disruptivity value. In particular, at threshold = 0.7, the SA rate reaches 98%, but the FA rate is also as high as 21%. At the opposite end of the spectrum, increasing the DPRF threshold to 0.95 causes the FA rate to decrease to 1.8%, with a strong reduction in the SA rate as well (60%). The best tradeoff, as corroborated also by the analysis of the offline data reported in section 2, is a DPRF threshold of 0.8, which guarantees SA > 90% and FA < 10%. This translates into 46 disruptive shots successfully predicted out of the 50 high-density disruptions, and 45 false alarm shots out of the 456 non-disruptive discharges.

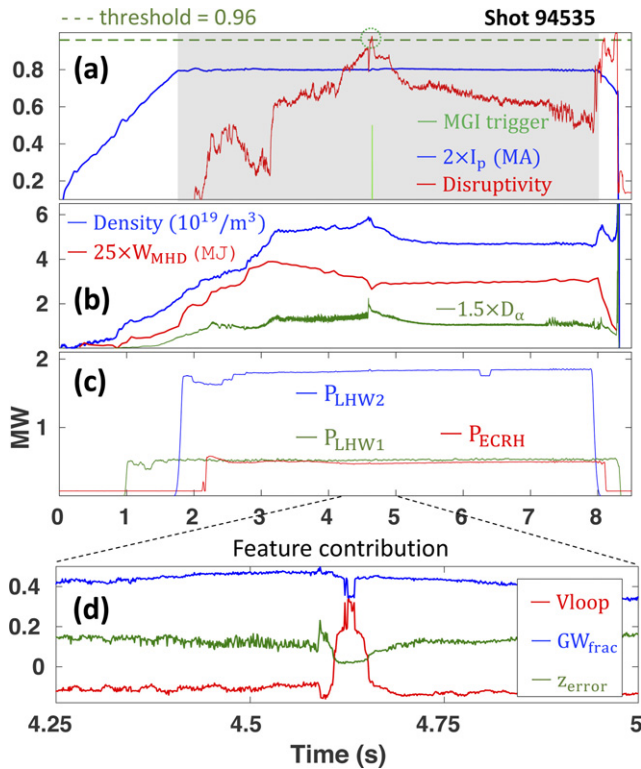
Having defined a DPRF warning threshold of 0.8, the cumulative fraction of disruptions detected as a function of the alarm time is studied, as shown in figure 9 and also detailed in other literature papers on disruption prediction [12,20,31]. This plot was obtained by analyzing the 50 high-density disruptive shots and recording the DPRF alarm time with a threshold of 0.8, with respect to the final disruption occurrence. The orange vertical dashed line at 30 ms defines the boundary of late detections, plotted to the left as a shadowed gray area. It is possible to see that almost all the correctly predicted disruptions (SA  $\sim 90\%$ ) occur at or before  $T_{\text{alarm}}$  of 40 ms, while just one disruption can be categorized as a late detection.

When it comes to premature detections, figure 9 shows that 28% of all disruptions are detected at  $T_{\text{alarm}} > \tau_{\text{class}}$  (1.7 s), which might likely be intended as an error bar or an uncertainty margin on the  $\tau_{\text{class}}$  choice. Since choosing  $\tau_{\text{class}}$  to be unique for training high-density shots is not an optimal choice, we expect to further reduce premature detections through a shot-by-shot selection of  $\tau_{\text{class}}$ .

The false alarms have been studied in detail as much as possible on a shot-by-shot basis to investigate the individual causes of such misclassifications. It is found that 15.5% of false alarms are caused by H–L back-transitions, with dynamics similar to those shown in figure 8, section 3.3. For the remaining misclassifications, an additional fraction (29%) shows a real-time density signal with values mistakenly higher than the actual plasma density (as registered with different offline diagnostics). As a result, the real-time calculated Greenwald density fraction is also higher than its real value, causing the Greenwald fraction contribution to the disruptivity to be higher than it should be, which eventually leads to false alarms. An example of this is shown in figure 10. In figure 10(b) the HCN density is compared with the density measured by the POLarimeter–INTERferometer (POINT) system [32]. The HCN density is obviously higher than the electron density measured by the POINT diagnostic, which is often caused by a power reduction of the HCN laser or a fast change of density, causing the  $GW_{\text{frac}}$  contribution to remain high, and eventually the disruptivity rises above 0.8 at  $t = 7.24$  s, thus triggering a false alarm. Unfortunately the POINT diagnostic is not routinely available in EAST, but



**Figure 7.** For discharges 94538 and 94537, (a) and (c) panels show the ion saturation current density contours measured by Langmuir probes located at the lower outer divertor, while the dashed lines in panels (b) and (d) show the eddy currents scaled by a factor of 10, as measured by the Rogowski loop sensor located at the lower divertor support. The vertical axis of the ion saturation current density contours refers to  $D_{\text{lower-divertor}}$ , the distance to the lower outer divertor corner. In (b) and (d) the dotted vertical lines are the disruption times of the two shots. The gray area is the time window when the MGI valve is turned on and neon gas is injected for shot 94538.



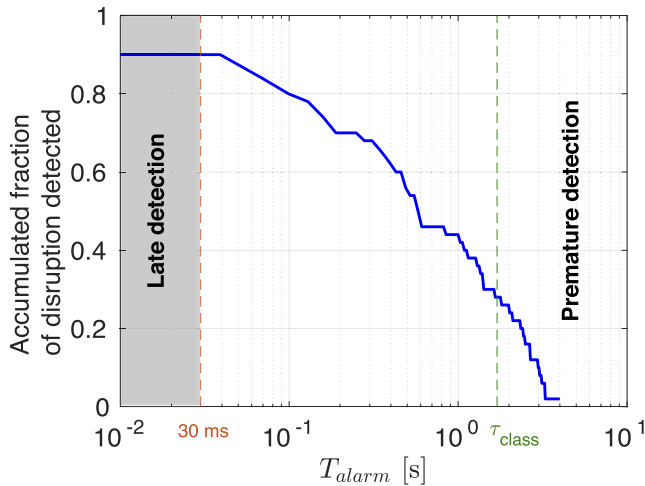
**Figure 8.** MGI triggered by a DPRF false alarm in discharge 94535: (a) disruptivity in red,  $I_p$  in blue scaled by a factor 2, and the MGI trigger signal (green spike) at DPRF threshold = 0.96 (for more than 10 ms). The gray shaded area highlights the flat-top phase. (b) Plasma density, plasma stored energy scaled by a factor of 25 and the  $D_\alpha$  signal scaled by a factor of 1.5. (c) Plasma heating power, including lower hybrid heating of two sources and ECRH power. (d) Three main feature contributions:  $V_{\text{loop}}$ ,  $GW_{\text{frac}}$  and  $Z_{\text{error}}$  near the time of MGI trigger signal.

**Table 3.** DPRF performance analysis for real-time discharges, scanning different thresholds for the disruptivity from 0.7 to 0.95. Here, ‘high-density disrupt shots’ means shots that disrupted with  $(GW_{\text{frac}})_{\text{max}} > 0.8$  during flat-top phase.

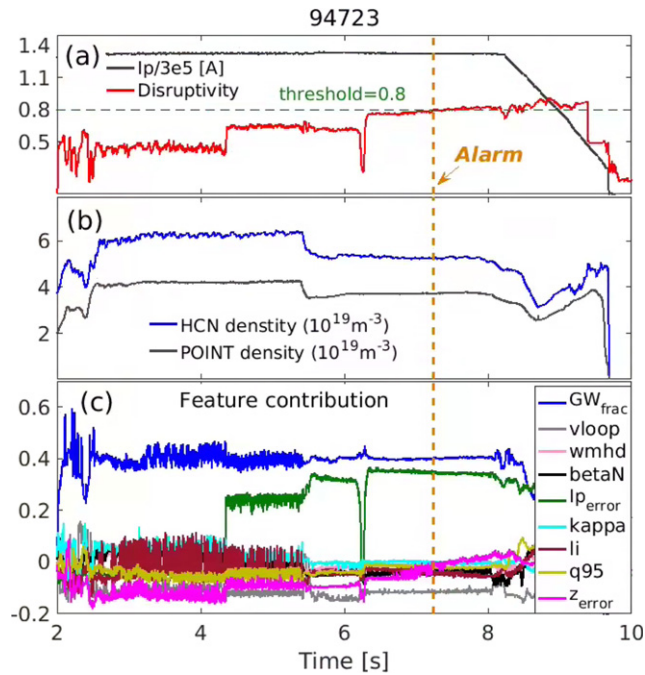
Warning duration (10 ms)	High-density disrupt shots (50)			Non-disrupt shots (456)
	Alarm threshold	SA rate	Mean alarm time (s)	
	0.7	0.98	1.5	0.21
	0.75	0.92	1.3	0.14
	0.8	0.92	1.2	0.099
	0.85	0.88	1.1	0.068
	0.9	0.76	0.62	0.037
	0.95	0.60	0.46	0.018

for those discharges with available data we recalculated the Greenwald density fraction and recomputed the disruptivity with the same model data: the recalculated offline disruptivity shows that the majority of these discharges would not have been classified as false alarms. However, the lack of POINT measurements on most false alarms hinders a robust statistical analysis, and therefore we can only speculate that the high fraction of false alarms—higher than the 29% we were able to fully investigate—is related to such problems in the correct density signal.

Excluding the two clear false alarm dynamics (H–L transitions and those related to an unphysical density signal), the remaining false alarm discharges still need more detailed investigations on aspects of signal disturbance, plasma operation, equilibrium reconstruction and so on in order to properly identify their causes. However, the final ~10% of the false alarm rate reported in table 3 represents a plateau of DPRF



**Figure 9.** Accumulated fraction of disruptions detected by DPRF among 50 high-density disrupt shots as a function of the alarm time (log scale) when setting the threshold to be 0.8. The gray-shaded area bounded by the 30 ms vertical dashed line is defined as a late detection area.



**Figure 10.** (a) Plasma current and disruptivity for non-disruptive discharge 94723 (false alarm). (b) HCN density and POINT density. (c) Feature area contribution of each input signal.

performances consistent with previous independent offline investigations [12].

## 5. Summary and future plan

In this paper, we discussed the importance of adopting data-driven algorithms to detect and prevent disruptions in real time on tokamak devices. These tools are essential to accelerate progress in disruption research and even more valuable when their interpretability is preserved to aid physics-based strategies [33]. In particular, we reported on recent progress

done to develop and use an algorithm based on random forests (DPRF) tailored onto high-density disruptions in EAST. During the 2019–2020 experimental campaign, DPRF ran in the background in EAST’s PCS but was also used in real-time mitigation experiments. The former allowed us to gather enough statistics to discuss DPRF’s overall performances, while the latter confirmed the viability of DPRF to be used as a trigger for the mitigation system when high-density disruptions occur. It was the first time that a machine learning-based disruption predictor was integrated and tested in the EAST control system.

DPRF is trained with nine zero-dimensional scalar signals, routinely available also in the EAST PCS, and provides as an output not only the probability of an impending disruption, i.e. the disruptivity, but also the so-called feature contributions, i.e. some explainability estimates to interpret in real time the drivers of the disruptivity. The real-time performances of DPRF were assessed also through the average computing time of the algorithm during PCS cycle times of 1 ms: DPRF, and the disruption prediction category, ran with an average computing time of 200–250  $\mu$ s, which satisfies the requirements for the real-time disruption warning. When the disruptivity signal increases up to a preset, configurable threshold for more than 10 ms, a trigger signal is sent to the MGI system. In section 3 we discussed the experimental results of mitigation experiments: when the MGI is triggered by the disruption warning signal, a massive amount of neon gas is injected, and the ion saturation current density at the lower divertor as well as eddy currents at the lower divertor support are successfully and greatly reduced, thanks to the DPRF trigger.

However, DPRF is not a perfect classifier for incoming disruptions: during experiments when the density constantly ramps up, the threshold power of H-mode confinement increases as well, but if the plasma heating power remains unchanged the plasma is likely to experience a degradation of confinement and a backward transition to the L-mode. This sudden transition from H- to L-mode was found to be a cause of DPRF false alarms, which reveals a present limitation of DPRF. Additionally, for a high fraction of false alarms the real-time density signal used was found to be mistakenly higher than the actual electron density measured with a different offline diagnostic. This caused higher estimates for the Greenwald density fraction, which in turn were driving the disruptivity signal above the trigger threshold.

Overall, DPRF has run in EAST’s PCS in the background for more than 1000 discharges, demonstrating its potential capability to be used as a real-time disruption predictor in EAST. A post-mortem analysis of these discharges reported in section 4 showed that an optimal warning threshold on the disruptivity should be kept at 0.8, such that a successful alarm rate of 92% can be reached (almost all of the disruptions are detected at or before 40 ms before the disruption), with a false alarm rate of 9.9%. Though the false alarm rate is still relatively high, this could be potentially reduced if better or more signals could be used. Because its first implementation in the EAST PCS was limited to only use nine scalar signals routinely available and relevant to high-density disruptions, as a next step, we plan to expand the input features available in



real-time to DPRF to include signals that better characterize the disruption dynamics of interest. For example, we plan to take advantage of the real-time digitization of a number of AXUV channels for future experimental campaigns: this would allow us to diagnose the evolution in time of the radiation profile in the core versus the edge regions of the plasma (following [18]), which could potentially help us improve the precision of DPRF. Beyond that, if DPRF had the capability to also detect H–L back-transitions or be integrated with warnings from hardware failures, it would help to avoid false alarms in future experiments. Additionally, we might want to focus on different disruption dynamics: as discussed in [34], the majority of EAST's disruptions are due to impurity accumulation events. Therefore, a real-time capable algorithm tailored to detect those cases would be highly beneficial to reduce the disruption rate in EAST and improve the experimental performances. Lastly, the choice of a refined  $\tau_{\text{class}}$  for each disruptive discharge, as opposed to the assumption of a unique value, was seen to increase the predictive performances of data-driven algorithms [31,35]. Therefore, as part of our future work, we plan to refine the  $\tau_{\text{class}}$  definition on a shot-by-shot basis by taking advantage of automated classification methods to extract metadata information that have been developed in the meanwhile [36].

## Acknowledgments

This material is based upon work supported by the National MCF Energy R&D Program of China, Grant No. 2018YFE0302100, and by the National Natural Science Foundation of China (Grant Nos. 12005264, 12075285, U1867222, 11875293 and 11775266). It was also partly supported by the US Department of Energy, Office of Science, Office of Fusion Energy Sciences, under Awards DE-FC02-04ER54698, DE-SC0014264 and DE-AC02-09CH11466, and by the President Foundation of Hefei Institutes of Physical Science, Chinese Academy of Sciences (Grant No. YZJJ2020QN11). Part of the data analysis was performed using the OMFIT integrated modeling framework [37]. One of the authors, Wenhui Hu, would like to express sincere thanks to Dr. Xingwei Zheng, Dr. Shouxin Wang, Dr. Xiang Gu and Dr. Qingquan Yang for their support and valuable discussion.

## ORCID iDs

W.H. Hu  <https://orcid.org/0000-0003-3420-2607>  
 C. Rea  <https://orcid.org/0000-0002-9948-2649>  
 D.L. Chen  <https://orcid.org/0000-0001-7093-3154>  
 Y. Zhang  <https://orcid.org/0000-0002-9510-4763>  
 J.C. Xu  <https://orcid.org/0000-0001-5886-3114>  
 K.J. Montes  <https://orcid.org/0000-0002-0762-3708>  
 L. Zeng  <https://orcid.org/0000-0003-4968-1401>

## References

- [1] Greenwald M., Terry J.L., Wolfe S.M., Ejima S., Bell M.G., Kaye S.M. and Neilson G.H. 1988 *Nucl. Fusion* **28** 2199
- [2] Greenwald M. 2002 *Plasma Phys. Control. Fusion* **44** R27
- [3] Hender T.C. et al 2007 *Nucl. Fusion* **47** S128–202
- [4] Mukhovatov V. et al 2003 *Plasma Phys. Control. Fusion* **45** A235
- [5] Maraschek M. et al 2018 *Plasma Phys. Control. Fusion* **60** 014047
- [6] Berkery J.W., Sabbagh S.A., Bell R.E., Gerhardt S.P. and LeBlanc B.P. 2017 *Phys. Plasmas* **24** 056103
- [7] Sengupta A. and Ranjan P. 2001 *Nucl. Fusion* **41** 487
- [8] Wang S.Y., Chen Z.Y., Huang D.W., Tong R.H., Yan W., Wei Y.N., Ma T.K., Zhang M. and Zhuang G. 2016 *Plasma Phys. Control. Fusion* **58** 055014
- [9] Zheng W. et al 2018 *Nucl. Fusion* **58** 056016
- [10] Tinguely R.A., Montes K.J., Rea C., Sweeney R. and Granetz R.S. 2019 *Plasma Phys. Control. Fusion* **61** 095009
- [11] Rea C., Granetz R.S., Montes K., Tinguely R.A., Eidietis N., Hanson J.M. and Sammuli B. 2018 *Plasma Phys. Control. Fusion* **60** 084004
- [12] Montes K.J. et al 2019 *Nucl. Fusion* **59** 096015
- [13] Rea C. et al 2021 Disruption prevention via interpretable data-driven algorithms on DIII-D and EAST 2020 *IAEA Fusion Energy Conf. EX/P1–25* (Nice)
- [14] Breiman L. 2001 *Mach. Learn.* **45** 5–32
- [15] Barr J.L. et al 2021 Development and experimental qualification of novel disruption prevention techniques on DIII-D 2020 *IAEA Fusion Energy Conf. EX/5-TH/6* (Nice)
- [16] Qiang X., Xiang G., Yinxian J., Haiqing L., Nan S., Yongfei C. and Xingde T. 2008 *Plasma Sci. Technol.* **10** 519
- [17] Gong X. et al 2019 *Nucl. Fusion* **59** 086030
- [18] Rea C., Montes K.J., Pau A., Granetz R.S. and Sauter O. 2020 *Fusion Sci. Technol.* **76** 912–24
- [19] Chawla N.V., Bowyer K.W., Hall L.O. and Kegelmeyer W.P. 2002 *J. Artif. Intell. Res.* **16** 321–57
- [20] Rea C., Montes K.J., Erickson K.G., Granetz R.S. and Tinguely R.A. 2019 *Nucl. Fusion* **59** 096016
- [21] Rea C. and Granetz R.S. 2018 *Fusion Sci. Technol.* **74** 89–100
- [22] Palczewska A., Palczewski J., Marchese Robinson R. and Neagu D. 2014 *Interpreting Random Forest Classification Models Using a Feature Contribution Method (Advances in Intelligent Systems and Computing vol 263)* (Cham: Springer) pp 193–218
- [23] Yuan Q.P., Xiao B.J., Penafior B.G., Piglowski D.A., Liu L.Z., Johnson R.D., Walker M.L. and Humphreys D.A. 2010 *Fusion Eng. Des.* **85** 474–7
- [24] Yuan Q.P. et al 2013 *Nucl. Fusion* **53** 043009
- [25] Huang Y., Xiao B.J., Luo Z.P., Yuan Q.P., Pei X.F. and Yue X.N. 2016 *Fusion Eng. Des.* **112** 1019–24
- [26] Huang Y. et al 2020 *Nucl. Fusion* **60** 076023
- [27] Chen D.L. et al 2020 *Plasma Phys. Control. Fusion* **62** 095019
- [28] Xu G. et al 2019 *Phys. Rev. Lett.* **122** 255001
- [29] Liu Z.X., Gao X., Zhang W.Y., Li J.G., Gong X.Z., Jie Y.X., Zhang S.B., Zeng L., Shi N. et al (EAST Team) 2012 *Plasma Phys. Control. Fusion* **54** 085005
- [30] Righi E. et al 1998 *Plasma Phys. Control. Fusion* **40** 721
- [31] Pau A., Fanni A., Carcangiu S., Cannas B., Sias G., Murari A. and Rimini F. 2019 *Nucl. Fusion* **59** 106017
- [32] Liu H.Q. et al 2016 *Rev. Sci. Instrum.* **87** 11D903
- [33] Piccione A., Berkery J.W., Sabbagh S.A. and Andreopoulos Y. 2020 *Nucl. Fusion* **60** 046033
- [34] Guo B.H. et al 2021 *Plasma Phys. Control. Fusion* **63** 025008
- [35] Aymerich E., Fanni A., Sias G., Carcangiu S., Cannas B., Murari A. and Pau A. (JET Contributors) 2021 *Nucl. Fusion* **61** 036013
- [36] Montes K.J., Rea C., Tinguely R.A., Sweeney R., Zhu J. and Granetz R.S. 2021 *Nucl. Fusion* **61** 026022
- [37] Meneghini O. et al 2015 *Nucl. Fusion* **55** 083008