Real-time prediction of high density EAST disruptions using Random Forest

W.H. Hu1, C Rea2, Q.P. Yuan1, K.G. Erickson2, D.L. Chen1, B Shen1, Y. Huang1, J.Y. Xiao4, J.J. Chen4, Y.M. Duan1, Y. Zhang1, H.D. Zhuang1, J.C. Xu5, K J Montes2, R S Granetz2, L. Zeng1, J.P. Qian1, B.J. Xiao1, J.G. Li1, and EAST Team

1 Institute of plasma physics, Chinese Academy of Science, Hefei China
2 Massachusetts Institute of Technology, Plasma Science and Fusion Center, Cambridge, MA USA
3 Princeton Plasma Physics Laboratory, Princeton, NJ USA
4 School of Nuclear Science and Technology, University of Science and Technology of China, Hefei China
5 School of Mechanical Engineering, Anhui University of Science and Technology, Huainan China

January 2021

Plasma Science and Fusion Center
Massachusetts Institute of Technology
Cambridge MA 02139 USA

This material is based upon work supported by the National MCF Energy R&D Program of China, Grant No. 2018YFE0302100. And it is supported by the National Natural Science Foundation of China (Grant No: 12005264, 12075285, U1867222, 11875293, 11775266). Additionally, it was partly supported by the U.S. Department of Energy, Office of Science, Office of Fusion Energy Sciences, under Awards DE-FC02- 04ER54698, DE-SC0014264 and DE-AC02-09CH11466. And it is partly supported by the President Foundation of Hefei Institutes of Physical Science of Chinese Academy of Sciences (Grant No: YZJJ2020QN11). Reproduction, translation, publication, use and disposal, in whole or in part, by or for the United States government is permitted.

Submitted to Nuclear Fusion
Real-time prediction of high density EAST disruptions using Random Forest

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

\textsuperscript{1} Institute of plasma physics, Chinese Academy of Science, Hefei China
\textsuperscript{2} Massachusetts Institute of Technology, Plasma Science and Fusion Center, Cambridge, MA USA
\textsuperscript{3} Princeton Plasma Physics Laboratory, Princeton, NJ USA
\textsuperscript{4} School of Nuclear Science and Technology, University of Science and Technology of China, Hefei China
\textsuperscript{5} School of Mechanical Engineering, Anhui University of Science and Technology, Huainan China

E-mail: crea@mit.edu

9 March 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 on 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 any disruption damage. DPRF’s average computing time of $\sim250$ $\mu$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 set of disruptions in which the electron density reached at least the 80% of the Greenwald density limit, using zero-dimensional signal routinely available to EAST PCS. Through offline analysis, an optimal warning threshold on 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 ($\sim15\%$) of the misclassifications are due to sudden transitions of plasma confinement from H-to L-mode, which often occur during high density discharges on EAST. By analyzing DPRF feature contributions, it emerges that the loop voltage signal is the 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
1. Introduction

One of the most critical parameters that affect high fusion performance in tokamak reactors is the plasma density. Empirical scalings from experimental observations across various devices show that when the plasma density increases up to near Greenwald density (\(n_{GW}\)) 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, 2, 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 to 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 more 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 (MLP), i.e. a fully connected feed-forward neural network model, was originally used in early 2000s to train a density limit disruption predictor. The authors collected 2000 samples from 23 discharges to train the predictor and a 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 (\(T_{alarm} \sim 30\) 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 to the availability of more efficient computational resources and the exploitation of larger sets of training data, which allows for further optimization of data-driven algorithms’ results, i.e. higher successful prediction rate and lower false prediction rate.
Random forest models have already been used to investigate possible disruption prediction solutions on DIII-D and Alcator C-Mod [11], reaching 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 labelled 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 on EAST to find dynamics similar to increasing plasma density plus 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 are considering high density disruptions that do 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 on EAST (DPRF) based on the Random Forest algorithm [14].

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 plasma control system (PCS) at 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 80% Greenwald density limit during the flattop phase of the plasma current. As a reminder, the Greenwald density limit is calculated as

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

(1)

where $I_p$ is the plasma current and $a$ is the minor radius. The plasma line-averaged density comes from the diagnostic of HCN interferometer [15]. 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 on EAST [12], which gathers numerous plasma signals sampled at 10 Hz during the flattop 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 input features. All training samples are therefore taken from the flattop phase of EAST discharges: a typical flattop duration is around 8 – 10 seconds, though some discharges have reached 100 seconds [16].

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 with [11], the samples between the beginning of the flattop phase and $t_D - \tau_{class}$ are regarded as ‘stable’ while samples between $t_D - \tau_{class}$ and $t_D$ are labelled as ‘unstable’. Here $t_D$ is disruption time and it is defined to be the time of $|dI_p/dt|_{\text{max}}$. $\tau_{class}$ is therefore defined with respect to the time of the disruption, $t_D$.

The choice of $\tau_{class}$ directly influences the performances of DPRF: $\tau_{class}$ 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 $F_\gamma$-score is adopted to evaluate DPRF’s performance on the validation set, which is calculated as:

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

(2)

Here precision=$TP/(TP+FP)$ and recall=$TP/(TP+FN)$, 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 classified as non-disruptive, and FN (False Negative) are disruptive shots that are misclassified as non-disruptive. $\gamma$ is chosen according to operational needs. For example, $F_\gamma$ = precision when $\gamma = 0$ and $F_\gamma$ = recall
Real-time prediction of high density EAST disruptions using Random Forest

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 $|dI_p/dt|_{\text{max}}$.

When $\gamma \to \infty$. 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 acting as training samples and the remaining subset acts as testing samples, or validation set; (ii) with every subset taking turns to be testing set random forest is trained repeatedly $K$ times for each value of $\tau_{\text{class}}$. Both $F_1$ and $F_2$ score are checked with $\gamma = 1$ or 2. The $F_2$ score privileges a higher accuracy on the disruptive class (the minority class), assigning a higher cost to those misclassifications. The $\tau_{\text{class}}$ was found to not differ too much between the $F_1$ and $F_2$ scores and the final optimal value of $\tau_{\text{class}}$ is found to be 1.7 s, with respect to the disruption time $t_D$. 

For the non-disruptive shots all of the flattop samples are collected and labelled 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 6 centralmost chords out of the available 64 ones. An updated definition of the $P_{\text{rad}}$ peaking that follows more closely what the authors have done on DIII-D [17] is already available offline and should be retroactively applied to EAST experimental database for disruption prediction studies. This is part of our next steps aimed at upgrading DPRF on EAST.
Real-time prediction of high density EAST disruptions using Random Forest

Figure 2: The probability histograms of the (a) Greenwald density fraction ($GW_{\text{frac}}$) (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. $P_{\text{rad}}$ peaking is not included in DPRF’s input features due to the unavailability of real-time digitization of radiation measurement, but its offline analysis support edge cooling mechanisms associated to density-limit disruptions.

In Figure 2 (a), around half of unstable samples are of $GW_{\text{frac}} > 0.8$, while most stable samples are of $GW_{\text{frac}} < 0.5$, which is in agreement with the data selection method mentioned above. The normalized internal inductance signal contains information on the current density profile distribution: the histograms in 2 (b) show a different behaviour for stable and unstable samples. The latter distribution is rather less peaked than the stable one, with a tendency to a bimodal behaviour 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 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 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 on 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 are collected, $\sim 16281$ of which are labelled as unstable. Although the training set is showing 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 [18]. 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 0-dimensional) plasma signals.
Table 1: DPRF real-time input signals and their description.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ip_{error}</td>
<td>(Ip − Ip_{target})/Ip_{target}</td>
</tr>
<tr>
<td>Vloop</td>
<td>Plasma loop voltage [V]</td>
</tr>
<tr>
<td>GW_{frac}</td>
<td>Greenwald fraction: ratio of plasma density to Greenwald density limit</td>
</tr>
<tr>
<td>β_{N}</td>
<td>Normalized ratio of plasma pressure to magnetic pressure</td>
</tr>
<tr>
<td>W_{MHD}</td>
<td>Plasma stored energy [J]</td>
</tr>
<tr>
<td>li</td>
<td>Plasma internal inductance</td>
</tr>
<tr>
<td>κ</td>
<td>Plasma elongation</td>
</tr>
<tr>
<td>q_{95}</td>
<td>Safety factor at 95% flux surface</td>
</tr>
<tr>
<td>z_{error}</td>
<td>(z − z_{target})/a, where z is the vertical position of plasma center, a is the minor radius of tokamak</td>
</tr>
</tbody>
</table>

and all of them are available in the plasma control system (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 models’ predictions, and thus reducing the model’s bias and variance. More methodological details about Random Forests can be also found in [11, 19].

As mentioned in Section 2.1, several critical model parameters (τ_{class}, disruptivity threshold, etc...) are found offline through a nested K-fold cross-validation approach to maximize the Fγ-score on a shot-by-shot basis, and through a parameters’ grid-search [12]. In particular, DPRF relies on a random forest composed of 920 decision trees (individual estimators) with a maximum depth of 15 layers – 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.

The training procedure is conducted similarly to what detailed in reference [11, 12] and [20]: the RandomForestClassifier core algorithm is taken off-the-shelf from the scikit-
learn open source library ‡ 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 treainterpreter package § 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 $\text{contrib}_m$ 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.

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

(3)

where $M$ is the total number of trees in the forest and the $n$ index runs on the number of the input features.

**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 alarms category. Disruption prediction, alarm and MGI system make up the close loop of disruption prediction and disruption mitigation system in EAST.

After training, the random forest model is translated from python into the C language and then compiled as an external library with the PCS so that it can run

‡ https://scikit-learn.org/stable/
§ https://github.com/andosa/treeinterpreter
Real-time prediction of high density EAST disruptions using Random Forest

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, Ip and Vloop are acquired from the magnetic diagnostics, the plasma density ne is acquired from density category where density is calculated with HCN diagnostic [21]. The plasma equilibrium and confinement parameters including li, WMHD, q95, βN, κ are provided by the ParaEquilibrium (PEFIT) category. The actual and target plasma vertical position z, ztarget and target plasma current Ip_target are from current and shape control category [22]. The calculation time of different categories over one cycle is different and the PEFIT category has the slowest calculation time. At EAST, the real-time equilibrium is reconstructed via parallelized GPU calculations [23, 24]. It takes 500 µs for PEFIT to complete one equilibrium iteration, parameter calculation and data transfer. Besides, DPRF calculates the disruptivity with an average computing time of ~200 µs – an example is shown in the mid panel of Figure 5. Therefore, the time cycle of disruption prediction category is set to be 1 ms, spending on signals reading, disruptivity calculation and result transfer to alarms category. If the disruptivity gets above a preset threshold (a parameter that can be configured during experiments) the alarms category starts to record it and if it lasts for 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’s experimental campaign, DPRF was tested while running using different warning thresholds for the disruptivity value, beyond the optimal 80% 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 Massive Gas Injection (MGI) system.

When DPRF runs in real-time experiments, it is usually enabled at the start of flattop phase because it is trained with all data taken from flattop 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 different situation. The disruptivity signal is sent to the alarm category in the PCS, which triggers the MGI valve for mitigation. 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 3 shots in Figure 4 are reported in Table 2: it shows that when the disruptivity threshold increases from 0.8 to 0.96, T_alarm decreases from 57 ms to 44 ms. Intuitively, a higher disruptivity threshold
Real-time prediction of high density EAST disruptions using Random Forest

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 threshold (a) 0.96 (b) 0.9 (c) 0.8.

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 80% for the disruptivity threshold as the optimal choice for maximizing DPRF performance and alarm time.

<table>
<thead>
<tr>
<th>shot number</th>
<th>threshold</th>
<th>$t_{\text{trigger}}$</th>
<th>$t_D$</th>
<th>$T_{\text{alarm}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>94520</td>
<td>0.80</td>
<td>5.708 s</td>
<td>5.765 s</td>
<td>57 ms</td>
</tr>
<tr>
<td>94521</td>
<td>0.90</td>
<td>7.995 s</td>
<td>8.048 s</td>
<td>53 ms</td>
</tr>
<tr>
<td>94522</td>
<td>0.96</td>
<td>8.040 s</td>
<td>8.084 s</td>
<td>44 ms</td>
</tr>
</tbody>
</table>

results in a shorter alarm time. This further validates the offline selection of 80% for the disruptivity threshold as the optimal choice for maximizing DPRF performance and the alarm time.

Figure 5 shows EAST shot 94520 during the last second before the disruption
Figure 5: Figure adapted from [13]: last 1 second 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 Greenwald fraction ($\text{GW}_{\text{frac}}$, $n/nG$ in the figure), $V_{\text{loop}}$, $I_{\text{error}}$.

In the middle panel of Figure 5(b) the DRPF computing time for each PCS cycle is shown to be around 200 $\mu$s, which is much shorter than the actual cycle time which is set to be 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 $I_{\text{error}}$ contribution also increases, signaling an increasing plasma resistivity and as a result, the disruptivity value increase to the preset threshold of 0.8. This behavior combined with a decreased contribution from the the Greenwald density fraction, hints at a disruption due to a radiative event or an impurity accumulation and not for a density-limit case. The increase of the disruptivity above the threshold for 10
ms leads anyways to trigger the MGI system valve. 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 to indicate a higher risk of disruption.

![Figure 6](image-url)

**Figure 6:** (a) The disruptivity and alarm threshold (dot line) (b) current and MGI trigger signal (c) density and (d) the total radiated power in bulk plasma evolution of two reference shots 94537 (blue line) and 94538 (red line). For shot 94538, 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 injection at around $t = 4.09$ s and immediately after that 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 plasma disrupts at $t = 5.24$ s.
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 in 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: this implied that the alarm was sent to the MGI system and triggered at around $t = 4.09$ s, injecting 4500Pa.L of Neon gas into the plasma.

We then analyze the mitigation effects comparing the two cases under three different aspects: (1) we compare the two discharges after the disruption, finding that the total radiation power in the bulk plasma is much smaller in 94538 than 94537. (2) At the same time in Figure 7(a), we show that the ion saturation current density is successfully reduced for the gas-injected case compared with Figure 7(c) case without gas injection. 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 [25] are greatly reduced in 94538 rather than 94537 – see Figure 7(b) versus (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 to the preset threshold of 0.96 for 10 ms an alarm is sent to the MGI system at $t_{\text{trigger}} = 4.65$ s, though without gas injection. Nevertheless, the plasma does not disrupt over the whole flattop 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 flattop phase, as it is trained with samples only from the flattop.

In this particular discharge (94535), the plasma transitions to H-mode near $t = 3$ s, when $D_\alpha$ signal shows that small edge-localized modes (ELMs) appear [26] (green curve Figure 8 (b)). With the density increasing, the plasma stored energy decreases until around $t = 4.65$ s, when a sudden decrease of $W_{\text{MHD}}$ happens indicating an H-L back transition. In EAST discharges, H-L back transition are empirically found to often occur during density ramping experiments. With density continuously ramping up, the threshold power of H mode increases as well and more energy is needed to sustain H mode [27, 28]. However, as shown in Figure 8(c) both the lower hybrid wave (LHW) and the electron cyclotron (ECRH) heating power remain almost constant while $W_{\text{MHD}}$ decreases constantly. At around $t = 4.7$ s, the $D_\alpha$ signal shows a peak and then 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
Real-time prediction of high density EAST disruptions using Random Forest

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) dotted vertical lines are the disruption times of two shots. The gray area is the time window when MGI valve is turning ON and Neon gas injected for shot 94538.

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, Vloop contributions increases suddenly, with a drop in $GW_{\text{frac}}$ and $z_{\text{error}}$ contributions. The great deviation in the increasing Vloop signal causes the disruptivity to spike and the DPRF false alarm. The loop voltage, usually signalling a change in the plasma resistivity, can be influenced not only by direct changes in the bulk plasma radiation and density but can also be affected by malfunctions of plasma heating/current driven system (hardware problem). Beside that, the Vloop signal might also be influenced by plasma radial motion due to a plasma’s dynamical response when $\beta_p$ suddenly changes during H-to-L transitions. All these causes should be taken into account when running a data-driven predictor, thus ensuring also its robustness against hardware failures. On EAST, we have empirically found that a variation of the Vloop signal may often not represent an incumbent disruption but rather be a cause of false alarms.
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 flattop 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.

4. Overall DPRF performance analysis during 2019-2020 EAST experimental campaigns

During the EAST experimental campaign that started from the Autumn of 2019 and ended in the Spring of 2020, DPRF was enabled in the PCS and kept to run in background during 1040 plasma discharges. In every one of these discharges DPRF was set to start working from $t = 2$ s, to be enabled during only the flattop phase. This allowed us to gather enough statistics to assess the overall performances of DPRF,
as shown in Table 3. Post-mortem analysis revealed that on a total of 1040 plasma discharges, 456 shots did not disrupt and could be categorized as healthy plasmas, therefore presenting a complete ramp up, flattop and ramp down phases. And 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 flattop phase of the plasma current. Discharges that matched the following conditions were excluded from performance evaluation: (i) shots with duration shorter than 2 s (DRPF 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 flattop phase.

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 $(GW_{frac})_{max} > 0.8$ during flattop phase.

<table>
<thead>
<tr>
<th>Warning duration</th>
<th>High density disrupt shots</th>
<th>Non-disrupt shots</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 ms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alarm threshold</td>
<td>SA rate</td>
<td>Mean alarm time [s]</td>
</tr>
<tr>
<td>0.7</td>
<td>0.98</td>
<td>1.5</td>
</tr>
<tr>
<td>0.75</td>
<td>0.92</td>
<td>1.3</td>
</tr>
<tr>
<td>0.8</td>
<td>0.92</td>
<td>1.2</td>
</tr>
<tr>
<td>0.85</td>
<td>0.88</td>
<td>1.1</td>
</tr>
<tr>
<td>0.9</td>
<td>0.76</td>
<td>0.62</td>
</tr>
<tr>
<td>0.95</td>
<td>0.60</td>
<td>0.46</td>
</tr>
</tbody>
</table>

The metrics that we defined to characterize DPRF performances are reported in the table. The successful alarm (SA) identifies the disruptive shot with a trigger signal sent out before the end of flattop phase. The false alarm (FA) identifies those cases when DPRF triggered a warning before the flattop end in non-disruptive discharges. SA and FA rate are normalized respectively by the number of disruptive and all non-disruptive shots.

By scanning 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 more stringent request for a higher disruptivity rate. In particular, at threshold = 0.7, SA rate reaches 98%, but FA rate is also as high as 21%. At the opposite end of the spectrum, increasing 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 on 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, shown in Figure 9 as detailed also in other literature papers on disruption prediction [12, 19, 29]. This plot was obtained by analysing 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. As it is possible to see, 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 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 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 $\tau_{\text{class}}$.

![Figure 9](image-url)

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

The false alarms have been studied in detail 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-to-L back transitions, with a dynamics similar to what shown in Figure...
Figure 10: (a) Plasma current and disruptivity for non-disruptive discharge 94422 (false alarm). (b) HCN density and POINT density. (c) Feature contribution of each input signal.

Additionally, 49% of false alarms are caused by the real-time density signal in the PCS, which is shown to have values mistakenly higher than the actual plasma density and this is often caused by the power reduction of HCN laser. As a result, the real-time calculated Greenwald density fraction is also higher than its real value, causing a higher Greenwald fraction contribution to the disruptivity than it should be, which eventually leads to false alarms. An example of it is shown in Figure 10. In Figure 10(b) the HCN density is compared with density measured by POlarimeter-INterferometer (POINT) system [30]. After around 2.5 s, the HCN density is obviously higher than POINT density, and the GW\textsubscript{frac} contribution also jumps up, and simultaneously disruptivity rises up to 0.8, which leads to a false alarm at 3.098 s. The remaining 35.5% of false alarms need more detailed investigation on aspects of signal disturbance, plasma operation, equilibrium reconstruction and so on, in order to identify their causes.
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 [31]. 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 background in EAST’s plasma control system 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 EAST control system.

DPRF is trained with nine zero-dimensional, scalar signals, routinely available also in 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. 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 µs, which satisfies the requirements for 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 have 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 plasma overall bulk radiation power, ion saturation current density at lower divertor and eddy currents at lower divertor support are successfully and greatly reduced, thanks to DPRF trigger.

However, DPRF is not a perfect classifier for incoming disruptions: during experiments when density constantly ramps up, the threshold of H-mode confinement increases as well, but if plasma heating power remains unchanged the threshold to sustain H-mode increases as well and the plasma is likely to experience a degradation of confinement and a backward transition to L-mode. This sudden transition from H- to L-mode was found to be a cause for DPRF false alarms, which reveals a present limitation of DPRF.

Overall, DPRF has run in EAST’s PCS in background during more than 1000 discharges, demonstrating its potential capability to be used as a real-time disruption predictor on EAST. A post-mortem analysis of these discharges reported in Section 4 has 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 more signals can be used. Because its first implementation in EAST PCS was limited to only use 9 scalar signals routinely available and relevant to high-density disruptions. As 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 campaign: 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 [17]), which could potentially help us improve precision of DPRF. Beyond that, if DPRF had the capability to also detect H-L back transitions or integrated with warning from hardware failures, it will help avoiding false alarm in future experiments. Additionally, we might want to focus also on different disruption dynamics: as discussed in [32], 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 disruptions rate on EAST and improve 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 [29, 33]. 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 [34].

Acknowledgments

This material is based upon work supported by the National MCF Energy R&D Program of China, Grant No. 2018YFE0302100. And it is supported by the National Natural Science Foundation of China (Grant No: 12005264, 12075285, U1867222, 11875293, 11775266). Additionally, it was partly supported by the U.S. Department of Energy, Office of Science, Office of Fusion Energy Sciences, under Awards DE-FC02-04ER54698, DE-SC0014264 and DE-AC02-09CH11466. And it is partly supported by the President Foundation of Hefei Institutes of Physical Science of Chinese Academy of Sciences (Grant No: YZJJ2020QN11). Part of the data analysis was performed using the OMFIT integrated modeling framework [35].

One of the authors Wenhui Hu would like to express sincere thanks to Dr. Shouxing Wang, Dr. Xiang Gu, and Dr. Qingquan Yang for valuable discussion.

References

REFERENCES


[18] N V Chawla K W Bowyer L O H W P K 2002 *Journal of Artificial Intelligence*


[20] Rea C and Granetz R S 2018 *Fusion Science and Technology* **74** 89–100
REFERENCES


