Interpretable data-driven disruption predictors to trigger avoidance and mitigation actuators on different tokamaks

Cristina Rea


1MIT PSFC, Cambridge, MA, USA
2ASIPP, Hefei, China
3General Atomics, San Diego, CA, USA
4GSFS, Univ. of Tokyo and Research Fellowships for Young Scientist, JSPS
5NFRI, Daejeon, Korea
6PPPL, Princeton, NJ, USA

1st IAEA Technical Meeting on Plasma Disruptions and their Mitigation July 20-23, 2020
Outline

• Disruption Prediction
  – Intro and motivations

• Overview Of Interpretable Algorithms Across Devices
  – DIII-D
  – EAST
  – KSTAR
  – JT-60U

• Summary And Conclusions
Plasma pushed close to operational limits often leads to instabilities onset or control faults: unintentional disruptions

- Disruptions related to **peak plasma performances**: higher stored energy, longer confinement times...

- **Real-time prediction** and **avoidance**, with **mitigation**, mandatory when scaling to reactor sizes and forces.

View from visible camera of disruption on Alcator C-Mod.
Statistical studies show complex chains of events: disruption precursors

Disruptions as final loss of control: successful precursors identification can inform plasma controllers on proper actuators.

De Vries et al. NF 51 (2011) 053018 “Survey of disruption causes at JET”
Active monitoring and prediction of soft/hard limits necessary to inform transition across ops boundaries


Proximity to stability boundaries need to be actively controlled: different control regimes
Interpretable ML models for disruption prediction useful resources to identify in real-time stability boundaries

- **On DIII-D and EAST**, the Disruption Prediction via Random Forest algorithm (DPRF) computes disruptivity and interprets its drivers in real-time.

- **On KSTAR**, similar exploration through Random Forest.

- **JT-60U**: Sparse Modeling by Exhaustive Search and Support Vector Machine.
Outline

• Disruption Prediction
  – Intro and motivations

• Overview Of Interpretable Algorithms Across Devices
  – DIII-D
  – EAST
  – KSTAR
  – JT-60U

• Summary And Conclusions
Deep Learning extracts general representations of disruptive behavior across devices

J.X. Zhu et al, “A new Deep Learning architecture for general disruption prediction across tokamaks”, this meeting

- Numerical experiments with aggregated DIII-D, C-Mod, and EAST data show DL learns disruptive characteristics: **device-independent knowledge**.

- Non disruptive data results **device-specific**, not improving performances.

- Limited disruptive data from target device still needed for prediction, as well as **all available non-disruptive** data.
Fusion Recurrent Neural Network (FRNN) with 0-D scalar inputs installed in DIII-D control system (PCS)

- FRNN Long Short-term Memory block implemented in DIII-D PCS.
  

- **Computing time < 2ms** for real-time eval.

- Associated actuator response studies in progress.

- FRNN **heat map** shows *disruption score at alarm time* most sensitive to **radiated core power** and **q95**.

W. Tang et al., accepted 2020 IAEA FEC paper TH/7-1Ra

---

**FRNN Sensitivity Study – shot 164582**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>plasma current</td>
<td></td>
</tr>
<tr>
<td>plasma current error</td>
<td></td>
</tr>
<tr>
<td>internal inductance</td>
<td></td>
</tr>
<tr>
<td>Locked mode amplitude</td>
<td></td>
</tr>
<tr>
<td>stored energy</td>
<td></td>
</tr>
<tr>
<td>q95 safety factor</td>
<td></td>
</tr>
<tr>
<td>Normalized Beta</td>
<td></td>
</tr>
<tr>
<td>Plasma density</td>
<td></td>
</tr>
<tr>
<td>Radiated Power Core</td>
<td></td>
</tr>
<tr>
<td>Radiated Power Edge</td>
<td></td>
</tr>
</tbody>
</table>

**FRNN disruption score**

**Alarm threshold**

T [ms] from 0 to 4000
DPRF supervised binary classification algorithm: identify transition non disruptive – disruptive phases

- **Fixed time** for transition from safe to disruptive operational space.
- Training set thousands of discharges, **agnostic to disruption type**.
- **Offline cross-machine** investigation 0-D features (flattop data).

DPRF is based on the Random Forest ensemble algorithm → collection of decision trees: 🌳🌳🌳

**Provides metrics of interpretability.**

---

C. Rea et al., Nucl. Fusion 59 (2019)
DPRF 2.0: to detect earlier disruptive precursors, feature engineering and dimensionality reduction

- **1D/2D profile information** compressed into **peaking factors**.

- Profile diagnostics mapped onto flux surfaces or core / divertor regions.

**Peaking factors are interpretable, easy to calculate in real-time**

A. Pau et al., IEEE TPS, 46 (2018)
A. Pau et al., Nucl. Fusion 59 (2019)
DPRF 2.0: improved label classification by detecting transition between specific operational boundaries

- **First disruptive precursors manually identified** for hundreds of discharges → Transition into unstable operational space: scenario detection.

- ML algorithms: training composition can skew the sensitivity of the model towards certain scenarios.

- **Need for (automated) identification of disruption causes.**

K. Montes et al, “Accelerating Disruption Database Studies with Semi-Supervised Learning”, *this meeting*

S. Sabbagh et al, “Progress on Tokamak Disruption Event Characterization and Forecasting Research and Expansion to Real-Time Application”, *this meeting*
DIII-D DPRF 2.0 – peaking factors added to 0-D inputs
Feature contributions to explain disruptivity drivers

Decision paths in DPRF trees provide average measures of explainability by assigning (±) contributions to input features during inference.

(see example in additional slides)

Access to disruptivity drivers in real-time:
monitoring of unstable plasma features
DPRF 2.0 shows real-time feature contribution computation (~ 200 µs) and successful ONFR* integration.

Closed the loop in the PCS by triggering early rapid shutdown, MGI, and ECH.

Assessed peaking factors as relevant metrics in scenario ~ ITER baseline.


N. Eidietis et al., 2018 Nucl. Fusion 58 056023
In progress: include DPRF 2.0 in DIII-D proximity control architecture to regulate stability and avoidance.

Disruptivity as general proximity of current plasma state to unstable ops space.

Feature contributions can be mapped onto controllable plasma parameters to regulate stability.

\[
\Delta \kappa = f_{\text{danger}} \cdot f_{\kappa, \text{contrib}} \cdot \text{sign} \left( \frac{d\kappa}{dt} \right) \frac{\Delta \kappa_{\text{target}}}{\Delta f_{\kappa, \text{contrib}}}
\]

J. Barr, “Control Solutions Supporting Disruption Free Operation on DIII-D and EAST”, this meeting.
DPRF disruptivity analogous to current probability of membership to disruptive class

The disruptivity $P_D$ can be used to:

- Predict the future probability of plasma survival $S(t + \Delta t | t)$ \[1\] or
- Model the instantaneous hazard \[2,3\] $h = \frac{d \ln S}{dt}$ to be used as probability generator.

Hazard function modeling connects dynamical systems and risk-aware control design by probability generation

\[ Pr[T > t | X_0 = x] = \mathbb{E} \left\{ \exp \left( - \int_0^t d\tau h(X_\tau) \right) \right\} \quad \text{s.t.} \quad dX = a(X)dt + b(X)dW \]

- Dynamical system \((a, b)\) either by ML or first principles or a combination; plasma state \(x\).
- Dependence on future actuation makes future event probability conditional: control design.
- Hazard function directly corresponds to (probabilistically calibrated) operational boundaries.
- Underutilized approach: only tearing mode events analyzed (in DIII-D) to date.

KEJ Olofsson et al 2018 PPCF 60
KEJ Olofsson et al 2018 FED 146
Outline

• **Disruption Prediction**
  – Intro and motivations

• **Overview Of Interpretable Algorithms Across Devices**
  – DIII-D
    – EAST
    – KSTAR
    – JT-60U

• **Summary And Conclusions**
DPRF installed in EAST PCS: feature contributions and disruptivity calculated in real-time in < 200 µs

- DPRF trained using 400 high-density (ne/nG > 0.8) disruptions and 400 non-disruptive data.

- Tested in real-time on 172 disruptive and 456 non-disruptive discharges.

- Tested in **closed-loop to fire mitigation system.**
EAST DPRF: disruptivity threshold of 0.8 guarantees SA ~89% and FA ~9% and alarm > 1 s

- SA: successful alarm, disruption detected in advance;
- FA: false alarm, alarm triggered for non-disruptive discharge.
Development of data-driven disruption prediction system using random forest method in KSTAR

- **Object**
  Development of disruption prediction system based on data-driven machine-learning methodology using KSTAR database

- **Database**
  - Total 1054 disruption shots from 2015 to 2018 KSTAR campaign
  - Label (disruptive / non-disruptive) based on 40 ms prior to thermal quench (40 ms: required time to activate disruption mitigation system, such as MGI or SPI)
  - Dataset: $l_{p,\text{error}}$, $f_{GW}$, $\delta B_{LM}$, $Z_0$, $q_{95}$, $V_{\text{loop}}$, and $I_i$

- **Training result**
  - Random forest, binary classification
  - Confusion matrix:

<table>
<thead>
<tr>
<th>True label</th>
<th>Non-disruptive</th>
<th>Disruptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-disruptive</td>
<td>14766 (0.463)</td>
<td>2033 (0.064)</td>
</tr>
<tr>
<td>Disruptive</td>
<td>1533 (0.048)</td>
<td>13531 (0.425)</td>
</tr>
</tbody>
</table>

- **Accuracy**
  - Accuracy on non-disruptive class: 90.6%
  - Accuracy on disruptive class: 86.7%

J. Lee et al., private communications
High-beta disruption prediction in JT-60U through exhaustive search and SVM

- Feature extraction via Sparse Modeling → **K-sparse Exhaustive Search**

![Graph showing time change of PSR (solid line) and FAR (dashed line)]

**Prediction success rate**

**False alarm rate**

---

**Binary classification** through linear Support Vector Machine (SVM) to extract decision function for the boundary: 

\[ f(x) = w \cdot x + b \]

---


---

Top combination:

\[ \beta_p, q_{95}, \kappa, f_{GW}, T_i \]

Results in **PSR ~ 95%**, **FAR ~ 15%** at **30 ms** before the disruption.
High-beta disruption prediction in JT-60U through exhaustive search and SVM

- Decision function obtained by retraining the SVM, after taking the log of the training data:

\[ f_{\text{exp}}(x) = e^{7.45 \beta_p^{5.39} q_{95}^{-8.29} \kappa^{7.40} f_{GW}^{4.50} T_i^{-0.120}} \]

Decision function parametrized from top combination of features enables disruption likelihood estimate

---

T. Yokoyama et al., Data-driven study of high-beta disruption prediction in JT-60U using exhaustive search, AAPPS 2019

Outline

• Disruption Prediction
  – Intro and motivations

• Overview Of Interpretable Algorithms Across Devices
  – DIII-D
  – EAST
  – KSTAR
  – JT-60U

• Summary And Conclusions
More than 20 years of research in disruption prediction have produced voluminous literature

<table>
<thead>
<tr>
<th>Device</th>
<th>References (incomplete list)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADITYA</td>
<td>Sengupta and Ranjan 2000 NF 40</td>
</tr>
<tr>
<td></td>
<td>Sengupta and Ranjan 2001 NF 41</td>
</tr>
<tr>
<td>Alcator C-Mod</td>
<td>Rea et al 2018 PPCF 60</td>
</tr>
<tr>
<td></td>
<td>Montes et al 2019 NF 59</td>
</tr>
<tr>
<td></td>
<td>Tinguely et al 2019 PPCF 61</td>
</tr>
<tr>
<td>ASDEX-U</td>
<td>Pautasso et al 2002 NF 42</td>
</tr>
<tr>
<td></td>
<td>Windsor et al 2005 NF 45</td>
</tr>
<tr>
<td></td>
<td>Aledda et al 2015 FED 96-97</td>
</tr>
<tr>
<td>DIII-D</td>
<td>Wroblewski et al 1997 NF 37</td>
</tr>
<tr>
<td></td>
<td>Rea and Granetz 2018 FST 74</td>
</tr>
<tr>
<td></td>
<td>Rea et al 2018 PPCF 60</td>
</tr>
<tr>
<td>EAST</td>
<td>Montes et al 2019 NF 59</td>
</tr>
<tr>
<td>JET</td>
<td>Windsor et al 2005 NF 45</td>
</tr>
<tr>
<td></td>
<td>Cannas et al 2004 NF 44</td>
</tr>
<tr>
<td></td>
<td>Cannas et al 2007 FED 82</td>
</tr>
<tr>
<td></td>
<td>Murari et al 2008 NF 48</td>
</tr>
<tr>
<td>JT-60U</td>
<td>Yoshino 2003 NF 43</td>
</tr>
<tr>
<td></td>
<td>Yoshino 2005 NF 45</td>
</tr>
<tr>
<td></td>
<td>Yokoyama et al. 2019 FED 140</td>
</tr>
<tr>
<td>J-TEXT</td>
<td>Wang et al 2016 PPCF 58</td>
</tr>
<tr>
<td></td>
<td>Zheng et al 2018 NF 58</td>
</tr>
<tr>
<td>NSTX</td>
<td>Gerhardt et al 2013 PPCF 60</td>
</tr>
</tbody>
</table>
Data-driven predictors to be adopted as last line of defense for disruption mitigation but...

• **Interpretable output** combined with **control** algorithms can inform the PCS on **disruption precursors** and be employed in **avoidance** schemes.
  – Frameworks exist to extract plasma **future survival** → Tinguely et al. or **instantaneous hazard** (as probability generator) for instabilities → Olofsson et al.

• **DPRF** provides **explainable predictions** – tested on **C-Mod, EAST, DIII-D**:
  – Works as **real-time scenario detector** (DIII-D, EAST).
  – To be integrated with **proximity controller** for continuous avoidance (DIII-D).

• Analogous efforts ongoing at international facilities:
  – J. Lee and J. Kim @ KSTAR
  – T. Yokoyama @ JT-60U
  – A. Pau and others @ JET, TCV, AUG
  – G. Dong et al. @ DIII-D

• **Ongoing work to design predictor for ITER**:
  – Few ITER disruptions might still be needed to design effective data-driven solutions.
    → J.X. Zhu et al.
    → J. Kates-Harbeck et al.
Acknowledgement:
This work was supported, in part, by the U.S. Department of Energy under DE-FC02-04ER54698, DE-FC02-99ER54512, DE-SC0014264, DE-SC0010720, DE-SC0010492.

Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
Decision paths in DPRF trees provide local measures of explainability through information gain and loss.

Predictions for forest of $M$ trees can be decomposed in the $K$ contributions from each evaluated input feature:

$$F(x) = \frac{1}{M} \sum_{m=1}^{M} \text{bias}_m + \sum_{k=1}^{K} \left( \frac{1}{M} \sum_{m=1}^{M} \text{contrib}_m(x, k) \right)$$
DPRF 0-D scalar input features – DIII-D and EAST

**DIII-D**

- $B_{r}^{n=1}/B_{\phi}$
- Locked Mode proxy
- $q_{95}$
- $n/n_{G}$
- $(I_{p} - I_{prog})/I_{prog}$
- $\ell_{i}$
- $\beta_{p}$
- $V_{loop}$
- $W_{mhd}$
- $r_{HWHM}(T_{e})/a$
- $P_{rad}/P_{inp}$

**EAST**

- $n/n_{G}$
- $V_{loop}$
- $(I_{p} - I_{prog})/I_{prog}$
- $\ell_{i}$
- $q_{95}$
- $W_{mhd}$
- $(z_{cur} - z_{prog})/a$
- $\beta_{n}$
- $\kappa$