Resistive Wall Mode Stability Forecasting in NSTX through Balanced Random Forests and Counterfactual Explanations

A. Piccione[1], J.W. Berkery[2], S.A. Sabbagh[2], Y. Andreopoulos[1]

[1] *Department of Electronic and Electrical Engineering, University College London, London, WC1E 7JE, UK*

[2] *Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA*

Abstract

- The Disruption Event Characterization and Forecasting (DECAF) code is a comprehensive framework including physics modules of several causal events leading to disruptions in tokamak plasmas.
- A new Random Forest (RF) forecaster combined with under/over-sampling methods is tested to predict whether the Resistive Wall Mode (RWM) will go unstable or not for a set of 134 NSTX discharges.
- Counterfactual Explanations [**1**] are used to gain insight into the physics beyond RWM stability and explore situations that could be actionable in a tokamak control room.

I. Machine Learning in DECAF

- The ML-determined ideal no-wall limit has been seen to perform well on NSTX data as well as on a limited number of MAST discharges.
- Neural network calculations [**2**] and six other plasma parameters (Table **1**) inspired by the full and reduced kinetic models [**3**] are used as inputs into a larger RWM stability model.
- The $n = 1$ MHD signal is used to provide a better compromise between true positives and false positives, since the presence of strong low frequency MHD activity almost always precludes an RWM from concurrently going unstable.

DECAF alias	Symbol
Normalized beta	β_N
betaN no-wall limit	$\beta_{N,no-wall}^{n=1}$
betaN with-wall limit	$\beta_{N,with-wall}^{n=1}$
Average E cross B frequency inside pedestal	$\langle\omega_E\rangle$
Average ion collision frequency inside pedestal	$\langle\nu_{ii}\rangle$
Low frequency, odd n MHD	$n = 1$ MHD

Table 1: Signals used as inputs into the ML-RWM stability forecaster.

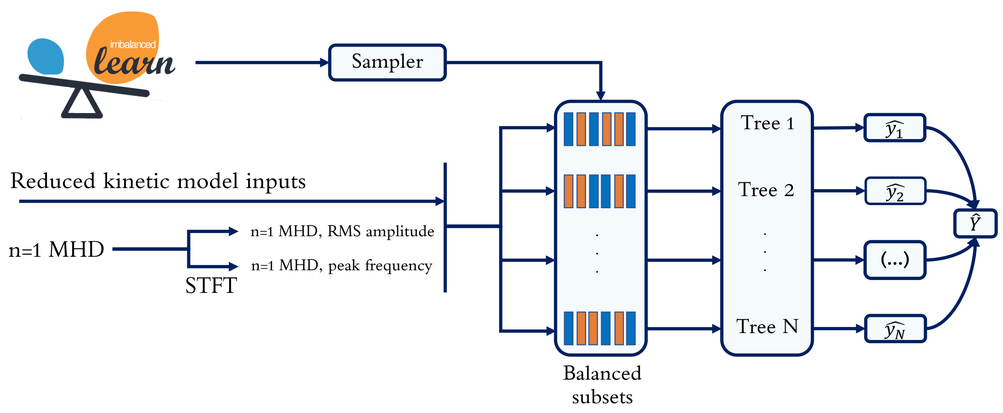
II. RF-RWM stability forecaster

BALANCED RANDOM FORESTS

An RF classifier is combined with sampling techniques to tackle imbalances in the dataset. Individual time points are resampled at a tree level in order to not lose any valuable information during training. Random Under-Sampling (BUSRF) gave the best compromise in terms of performance and training time ($\sim$ 3x faster than base RF).

OVERALL PERFORMANCE

Models are trained via stratified cross-validation using a hysteresis thresholding method [**4**] as warning criterion. 10/11 unstable RWMs were correctly identified (TPR = 90.9%), with 2/17 wrongly triggered stable RWMs (FPR = 11.7%) in test phase.



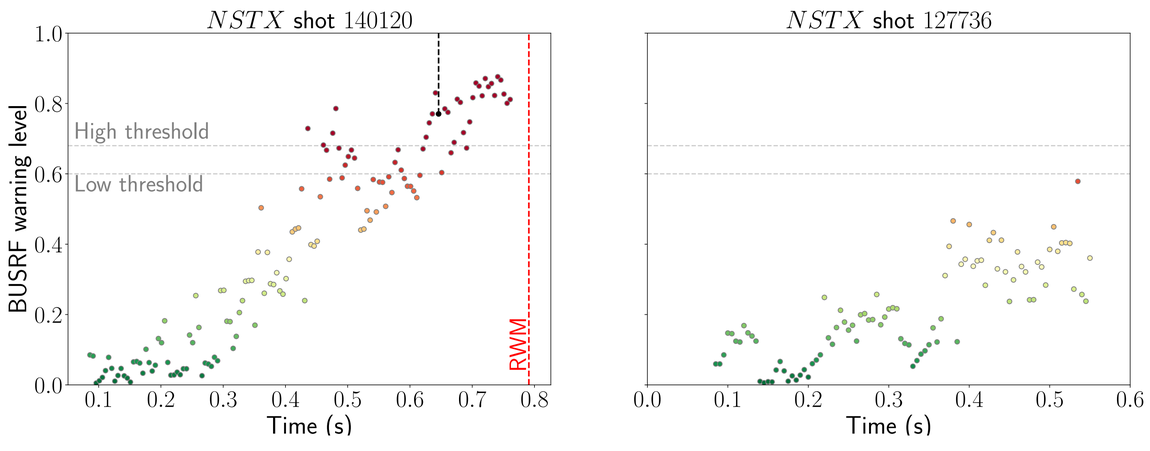


Fig. 1: Architecture of RWM's predictor pipeline (top). Comparison of RF's warning level for an unstable RWM (bottom left) and a stable one (bottom right).

III. Counterfactual Explanation of MHD activity

- The DiCE algorithm [**2**] generates hypothetical realities that contradict the observed facts by minimizing a loss function.

$$\mathbf{C}(\mathbf{x}) = \underset{c_1,\ldots,c_k}{\arg\min} \underbrace{\frac{1}{k}\sum_{i=1}^{k}\mathcal{L}(f(c_i), y)}_{\text{Hinge loss}} + \underbrace{\frac{\lambda_1}{k}\sum_{i=1}^{k}\mathcal{D}_1(c_i, \mathbf{x})}_{\text{Proximity}} - \underbrace{\lambda_2\mathcal{D}_2(c_1, ..., c_k)}_{\text{Diversity}} ,$$

- The first application aims to explore the effect of low frequency $n = 1$ MHD activity on the underlying model's predictions. DiCE is constrained inside bounds that are inspired by physics knowledge.

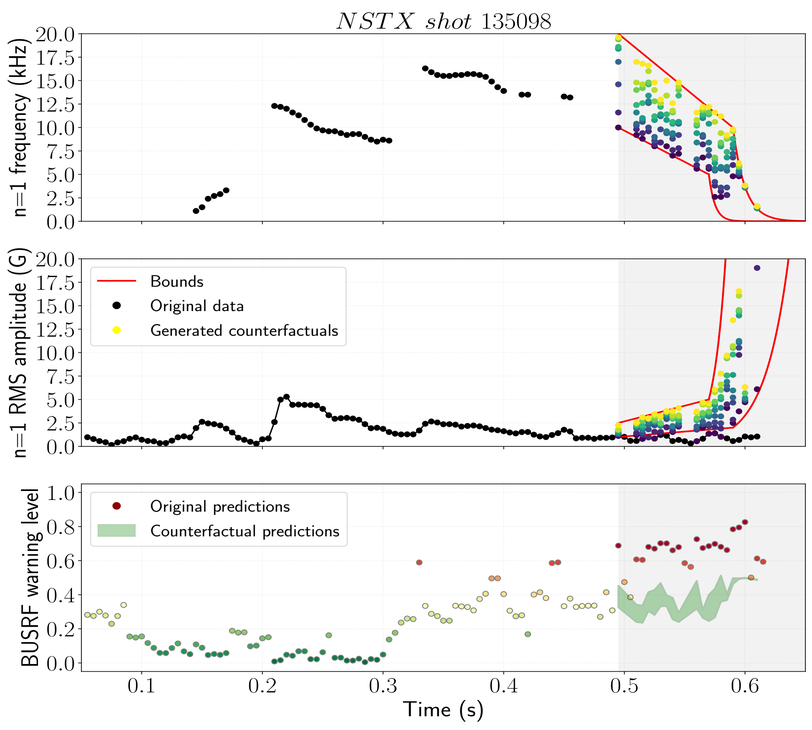


Fig. 2: MHD activity and DiCE-generated counterfactuals for NSTX shot 135098. The grey shaded area indicates the region with originally no MHD activity.

IV. Simulated β_N control

- DiCE is used to infer the highest β_N level that would have kept the plasma RWM stable. λ_2 is set to 0 since the value closest to the actual measurement is desired. This represents a potentially valuable input to complex control algorithms [**5**] already planned for NSTX-U operations.

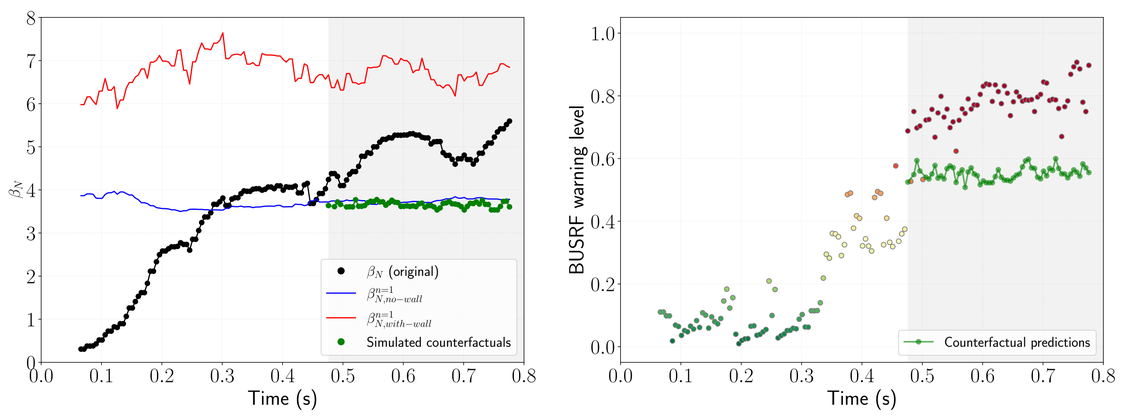


Fig. 3: DiCE-informed β_N levels to keep the warning level below the low threshold for NSTX shot 140123.

Conclusions and further development

Implementation of ML tools in DECAF continues with the introduction of a new RWM stability forecaster to support the existing reduced kinetic model. Counterfactuals have been used for the first time to explore actionable scenarios and the implementation of a more sophisticated β_N/rotation ML-based controller is under way. Future work also includes applications to similar tokamaks, such as MAST-U, since projections show a region of potential high-beta stability [**6**].

Acknowledgements/References

Supported by U.S. DoE under contract DE-SC0018623 (Columbia), and by EPSRC grants EP/R025290/1 and EP/P02243X/1 (UCL).

[1] R.K. Mothilal et al., Proceeding of the 2020 Conference on F.A.T. (2020)

[2] A. Piccione et al., Nuclear Fusion **60**, 046033 (2020)

[3] J.W. Berkery et al., Physics of Plasmas **24**, 056103 (2017)

[4] K.J. Montes et al., Nuclear Fusion **59**, 096015 (2019)

[5] M.D. Boyer et al., Nuclear Fusion **55**, 053033 (2015)

[6] J.W. Berkery et al., Plasma Phys. Control. Fusion **62**, 085007 (2020)