Predicting resistive wall mode stability in NSTX through balanced random forests and counterfactual explanations

OPEN ACCESS
IOP Publishing | International Atomic Energy Agency Nuclear Fusion

Nucl. Fusion **62** (2022) 036002 (17pp) https://doi.org/10.1088/1741-4326/ac44af

Predicting resistive wall mode stability in NSTX through balanced random forests and counterfactual explanations

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

[1] Department of Electronic and Electrical Engineering, University College London, WC1E 7JE, United Kingdom
[2] Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, United States of America

E-mail: a.piccione@ucl.ac.uk

Received 29 September 2021, revised 30 November 2021
Accepted for publication 20 December 2021
Published 18 January 2022

Abstract

Recent progress in the disruption event characterization and forecasting framework has shown that machine learning guided by physics theory can be easily implemented as a supporting tool for fast computations of ideal stability properties of spherical tokamak plasmas. In order to extend that idea, a customized random forest (RF) classifier that takes into account imbalances in the training data is hereby employed to predict resistive wall mode (RWM) stability for a set of high beta discharges from the NSTX spherical tokamak. More specifically, with this approach each tree in the forest is trained on samples that are balanced via a user-defined over/under-sampler. The proposed approach outperforms classical cost-sensitive methods for the problem at hand, in particular when used in conjunction with a random under-sampler, while also resulting in a threefold reduction in the training time. In order to further understand the model's decisions, a diverse set of counterfactual explanations based on determinantal point processes (DPP) is generated and evaluated. Via the use of DPP, the underlying RF model infers that the presence of hypothetical magnetohydrodynamic activity would have prevented the RWM from concurrently going unstable, which is a counterfactual that is indeed expected by prior physics knowledge. Given that this result emerges from the data-driven RF classifier and the use of counterfactuals without hand-crafted embedding of prior physics intuition, it motivates the usage of counterfactuals to simulate real-time control by generating the β_N levels that would have kept the RWM stable for a set of unstable discharges.

Keywords: machine learning, resistive wall mode, counterfactuals, random forest

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

1. Introduction

The resistive wall mode (RWM) is a global mode of instability of high pressure tokamak fusion plasmas that can lead to disruption of the plasma current and termination of the discharge [1, 2]. Since these disruptions can lead to physical damage to the machines, a method of early detection and forecasting of RWM stability is desired. In past post-discharge analysis efforts [3, 4], an exponential rise in an $n = 1$ toroidal mode number poloidal magnetic field measurement (known as the RWM sensor) on the time scale of magnetic flux penetration through conducting surfaces surrounding the plasma, τ_w, was used as the primary indicator of RWM instability. However, while this signal defines the existence of the RWM it requires that the mode be unstable. It is highly preferred to give an ear-

[*] Author to whom any correspondence should be addressed.

lier warning of the mode potentially becoming unstable, therefore other predictors such as the normalized beta, β_N, level and the lack of indication of a locked tearing mode (LTM) were also examined. A tearing mode is another mode of instability of tokamak plasmas that is more localized to rational magnetic surfaces rather than having a more global eigenfunction such as the RWM [5, 6]. The presence of low frequency rotating magnetohydrodynamic (MHD) activity, which can lead to an LTM, has been seen to almost always preclude an RWM from going unstable at the same time [3].

Analysis of the physics of RWM instabilities identified other important plasma characteristics such as plasma rotation and collisionality [4, 7], although the dependencies were too complex to handle via manual identification efforts. For example, a trained specialist could not simply look at a plasma rotation profile and use it as an indication of RWM stability or instability; what mattered were resonances between those rotation profiles and certain particle motions [8]. Complex physics software tools that analysed these kinetic resonances, such as MISK [9] and MARS-K [10] were developed and benchmarked [11, 12]. However, while these tools were useful to understand the physics of RWM stability, they were too complex to execute in real-time to predict instabilities in a manner such that they could be avoided. An alternative approach was then developed, where the physics of RWM instability uncovered by these codes was distilled into a reduced model that maintained the major physics, albeit in a more tractable form that could potentially be executed in real-time [13]. This reduced kinetic model was included in the disruption event characterization and forecasting (DECAF) code [13–16], which includes many other models of physical events that can predict potential disruptions in addition to RWM destabilization.

Recently, machine learning (ML) algorithms have been explored for disruption prediction and avoidance as well [17–21]. In one extreme, these could be completely ‘black box’ approaches where available data is indiscriminately fed in and used to train a disruption predictor. In other cases the models have been built in order to produce results that are less difficult to interpret and allow the gain of physics insight (i.e. feature contribution, stability maps etc) [22, 23]. An alternative is to use a physics-guided ML approach [24, 25]. In such cases, the known physics of the problem (i.e. domain-specific knowledge) is used to pre-process input data, to constrain the learning objective as well as to interpret the output. ML techniques have now been used for one piece of the kinetic RWM stability problem—determining the *ideal* stability [26], including a physics-guided framework [27], and this piece has been incorporated into DECAF.

This paper extends this work to use ML to predict whether the RWM will go unstable in NSTX. Specifically, a random forest (RF) based algorithm is tested on discharges where human analysis has determined the time of RWM instability, or lack thereof. Inputs to the algorithm include the previous physics-guided neural network (PGNN) determination of the ideal no-wall β_N limit [27], an expression for the with-wall limit, the measured β_N, two measured quantities related to the rotation and collisionality, and finally a signal indicating the presence of a rotating MHD mode. The latter signal is expected to provide a better compromise between true and false positives in the proposed approach and is also used in DECAF as part of an MHD/LTM warning module [15]. In DECAF the *absence* of an MHD warning is used in conjunction with a simple threshold test on the RWM sensor signal to indicate the possible presence of a growing RWM.

The method presented here serves four main purposes:

(a) To propose a different approach than the reduced kinetic model (RKM) [13] for RWM stability forecasting, which uses experimental data inspired by the full and reduced kinetic models without requiring the computation of kinetic and potential energies,

(b) To show that the RKM and the ML approach have the potential to cooperate towards a more reliable RWM stability forecaster since the latter extends the former’s domain of applicability,

(c) To prove that the combination of sampling approaches with tree-based classifiers helps in tackling strong imbalances in the dataset, an aspect of crucial importance since some events leading to disruptions occur with very low prevalence [28], and

(d) To propose a model-agnostic tool to interpret ML-based disruption predictions via counterfactual explanations and determinantal point processes (DPP) [29]. The exploration of *what if* conditions and how they affect the outcome has a twofold benefit. Firstly, it can assess the reliability of an ML algorithm by verifying that its counterfactual predictions agree with the known underlying physics. Secondly, one can use the counterfactual generation process to explore situations that are *actionable* in a real control system, such as lowering the β_N level slightly above the computed no-wall limit.

2. The reduced kinetic RWM model

Tokamak fusion plasmas are theoretically stable to global, ideal MHD instabilities up to the so-called no-wall beta limit, $\beta_{N,\text{no-wall}}^{n=1}$. Specifically, if no electrically conducting wall is present, the plasma is theoretically unstable above this limit to external kink-ballooning modes driven by the free energy of current or pressure gradients. Successful wall stabilization of kink/ballooning instabilities uncovered the reduced growth rate, yet still disruptive, RWM [2]. The RWM exists between the no-wall and the higher with-wall limit, $\beta_{N,\text{with-wall}}^{n=1}$, at which point not even wall effects can stabilize the ideal kink/ballooning modes. Stabilizing the RWM is of paramount importance since it grows on a timescale ($\tau_w \sim ms$) that is still fast compared to the duration of the plasma shot. Since magnetically confined plasma fusion devices need to operate at high β in order to improve plasma confinement efficiency, it is strongly desirable to operate above the no-wall limit. Past tokamak experiments found the fortuitous result that plasmas could be stably operated above the no-wall limit [6, 30] with either passive stabilization or active mode control [31]. Understanding the physics of passive stabilization is key to relying on it and projecting it to the operation of future devices, and theoretical attention turned to kinetic modifications of ideal MHD theory [32].

In the DECAF code, the RKM was implemented and performed well on a limited number of discharges from the NSTX tokamak [13]. In the RKM, the growth rate of the RWM was calculated using an energy principle approach [33], converting the force balance into an equation to determine mode stability based on a potential energy functional δW. The resulting dispersion relation for the complex mode frequency is:

$$(\gamma - i\omega_r)\tau_w = -\frac{\delta W^{n=1}_{\text{no-wall}} + \delta W_K}{\delta W^{n=1}_{\text{with-wall}} + \delta W_K}. \qquad (1)$$

Three terms needed to complete the RKM calculation are: the no-wall ideal term, $\delta W^{n=1}_{\text{no-wall}}$, the with-wall ideal term, $\delta W^{n=1}_{\text{with-wall}}$, and finally the more complex kinetic term δW_K. Rather than replicate the calculation of the RKM, the approach taken here is to use signals inspired by the physics model as inputs to an ML algorithm which will then return an RWM warning level. This can be considered an alternative to the growth rate from the RKM calculation.

2.1. No-wall and with-wall beta limits

The evaluation of the change in plasma potential energy due to a perturbation of the confining magnetic field without the presence of a conducting wall, $\delta W^{n=1}_{\text{no-wall}}$, was covered extensively in previous work [27]. More specifically, a RF regressor and a PGNN were tested on a large database of equilibria from the NSTX tokamak and used to reproduce the output of the DCON stability code [34] and to get an improved closed form equation for the no-wall limit. The physics guidance of the ideal MHD model in its extension outside the domain of the NSTX experimental data helped in carrying cross-device calculations to the MAST tokamak as well. Therefore, for the no-wall part of the equation, the quantity $\beta^{n=1}_{N,\text{no-wall}}$ from the previous PGNN [27] is used (see appendix).

Similarly to the no-wall limit, there is a higher with-wall beta limit used in the calculation of $\delta W^{n=1}_{\text{with-wall}}$ in the RKM. However, unlike the no-wall case, there does not presently exist a large database of calculations of $\delta W^{n=1}_{\text{with-wall}}$ or $\beta^{n=1}_{N,\text{with-wall}}$ to train a neural network on. Rather, a somewhat simple expression dependent on the 'pressure peaking factor', i.e. the ratio of central to average pressure, was found to be adequate for NSTX [13]. The same expression will be used in this work (see appendix), although it should be possible in the future to train an ML model to give $\beta^{n=1}_{N,\text{with-wall}}$ based on more plasma parameters.

2.2. Kinetic effects

The term δW_K results from a volume integration of the plasma displacement eigenfunction dotted with the divergence of the perturbed kinetic pressure tensor. The perturbed kinetic pressure tensor is found by taking a moment of the perturbed distribution function of the particles. In the end, the expression for δW_K is an integral over particle's energy, pitch angle, and magnetic surface of a fraction which includes the frequencies of the particle's motion including bounce, precession drift, $E \times B$ frequency (or plasma rotation), and collisionality [35]. When these frequencies match with opposite sign in the denominator, a resonance between particle motions is expected, which leads to a large fraction, a large δW_K, and a stabilizing effect. Physically, the stabilizing effect can be thought of as a match between particle motions and the mode, which then allows efficient transfer of energy from the potential growth of the mode to the particle motions, thus damping the growth. Consideration of the effect of collisions showed that reduction of collisionality will reduce the collisional dissipation that is important when plasma rotational resonances are not present, but conversely can also reduce the damping of resonant kinetic stabilizing effects, allowing them to be more powerful [7].

In order to reduce the complexity of this calculation for potential use as a real-time disruption warning, the RKM model implemented in DECAF [13] dispensed with the integration and instead used Gaussian functional forms for δW_K resonances with particle precession and bounce frequencies. These functional forms depend on the measured values of the $E \times B$ frequency and collisionality, $\langle\omega_E\rangle$ and $\langle\nu_{ii}\rangle$, both averaged inside the plasma pressure pedestal. These quantities are calculable in real-time if measurements such as plasma rotation, ion and electron density and temperature profiles are available. These profiles are also available in the DECAF database of analysed discharges and will be used in the present work as inputs to the described ML approach. The full expression for δW_k and the RKM can be found, for reference, in the appendix. However, neither of these expressions are used in the present work. Instead, the ML approach essentially creates its own use of the $\langle\omega_E\rangle$ and $\langle\nu_{ii}\rangle$ input features for predicting RWM stability.

Table 1. List of NSTX signals used as an input to the RF model. Left column provides the DECAF aliases as they appear in the database, while in the right column are the corresponding labels used throughout this paper.

DECAF alias	Symbol
Normalized beta	β_N
betaN no-wall limit	$\beta^{n=1}_{N,\text{no-wall}}$
betaN with-wall limit	$\beta^{n=1}_{N,\text{with-wall}}$
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	Odd-n MHD

3. The DECAF RWM database

3.1. Data considerations

In order to develop an ML-based RWM predictor, we have gathered a set of 134 NSTX discharges where a human validation step had been carried out to assess whether the primary DECAF indicator of RWM stability was in agreement with an expert's judgement. For 44 shots, the coupled analysis identified the time at which the RWM went unstable, while the remaining are experimentally RWM stable. Listed in table 1 are the aforementioned plasma parameters with the DECAF aliases and the corresponding symbols used later in this paper.

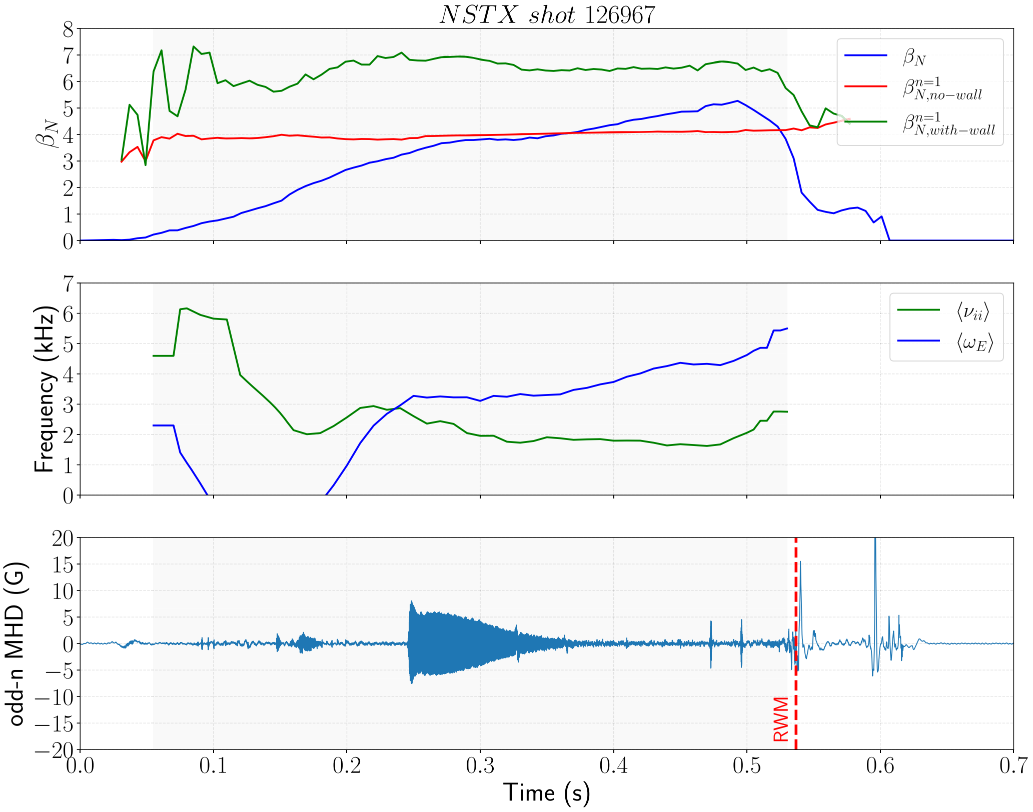

Figure 1. Features selected to train the RWM stability predictor in DECAF for NSTX shot 126967. The $\langle\omega_E\rangle$ and $\langle\nu_{ii}\rangle$ quantities have been causally smoothed for visualization purposes. Human analysis has identified an unstable RWM at $t = 0.537$ s, marked as a red vertical dashed line in the bottom frame. The shaded grey area represents the cutting window determined by the preprocessing step.

Table 2. Sub-categories identified in the NSTX database after performing the preprocessing step. The ones in which the no-wall limit is never crossed are kept in order to let the model be able to discern these cases from the more common ones.

Category	Description	Count
I	Above the no-wall limit	58 (stable) + 42 (unstable)
II	Never above the no-wall limit or crossed after disruption	27 (stable) + 1 (unstable)
III	No-wall limit crossed outside the cutting window	5 (stable) + 1 (unstable)

For this dataset, rotation profiles were available as a function of normalized poloidal magnetic flux (Ψ_N). However, we have decided to deploy a model based only on 0D signals since the required radial profiles might not be fully available in a future real-time system [13, 36]. Moreover, the usage of averaged quantities allows us to run ML tools in DECAF in parallel with the RKM without having to interact with a completely different set of plasma parameters. Additionally, two more features are obtained by processing a low frequency odd-*n* MHD signal currently used in DECAF to assess the presence of a rotating MHD mode, as shown later. Finally, another substantial difference between the RKM and the model developed here is that the latter uses β_N along with the no-wall and with-wall limits without combining them into the well-known $C_\beta = (\beta_N - \beta^{n=1}_{N,\text{no-wall}})/(\beta^{n=1}_{N,\text{with-wall}} - \beta^{n=1}_{N,\text{no-wall}})$ parameter [13]. In fact, since C_β is normally only computed in the range of $\beta^{n=1}_{N,\text{no-wall}} < \beta_N < \beta^{n=1}_{N,\text{with-wall}}$, we are in practice extending the domain of applicability of the DECAF RWM stability module.

3.2. Data preprocessing

A typical discharge in NSTX lasts up to 1.0–1.3 s and most of the measurements are available with a time resolution of 5 ms or less, hence the dataset is comprised of roughly 2×10^4 individual time slices. In this work, we have decided to not restrict our analysis to the flat-top phase or to the time window in which β_N is above the no-wall limit, but rather we cut the data to the range of times during which all the signals are available, followed by an interpolation on a common time scale of 5 ms. The resampling process never uses future information and it is done by looking at the previous and closest time point in the original timebase [19]. It is important to highlight that all the signals but $\langle\omega_E\rangle$ and $\langle\nu_{ii}\rangle$ are always available for the entire duration of the shot. In a few unstable cases (see figure 1), the latter are absent at the very end of the discharge

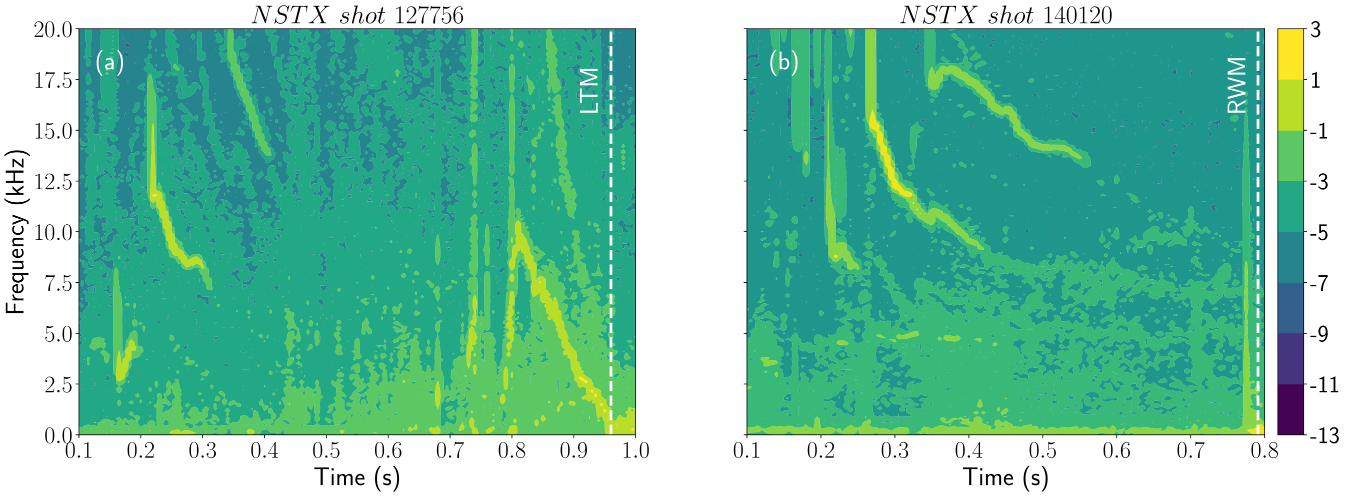

Figure 2. Low frequency odd-n MHD spectrum for (*a*) shot 127756, which ended in an LTM and for (*b*) the RWM-unstable shot 140120. The two different events are indicated by vertical white dashed lines. The colour coding shows how the peak frequency traces evolve differently in the two scenarios.

leading to cutting out the β_N collapse, which is a non-trivial indicator of a possible RWM instability. To address this issue, we initially considered to forward-fill the last valid point, however assuming that the collisionality and the $E \times B$ frequency do not change towards the end of the shot is not physically plausible. Fortunately, these few cases did not undermine the performance of our model.

Table 2 lists the three different types of discharges uncovered by our preliminary analysis. It is worth noticing that in 34 shots the no-wall limit is either never reached during the discharge, crossed after the disruption, or crossed before the disruption but outside the cutting window. Moreover, in just 2 of those 34 has the human analysis identified an unstable RWM. The choice of not ruling out low beta discharges is key from the physics side as well as from the ML point of view. In fact, although an unstable RWM below the computed no-wall limit is a rare event, a properly trained model should have access to all the possible cases.

With regards to the odd-n MHD signal, we have applied a short-time fourier transform to find mode peak frequency within the previous 5 ms. The root-mean-square (RMS) amplitude for the same time frame is also extracted using the librosa package. The usefulness of these two signals will be shown in detail in section 6.

Figure 2 displays the magnetic spectrogram for (*a*) an RWM-stable discharge in which the mode rotation slowly decreases and locks at $t = 0.96$ s, and for (*b*) a shot in which the RWM went unstable at $t = 0.79$ s. One can notice the substantial difference in the time evolution of the peak frequency, as well as in the amplitude level (in $\log_{10}$-scale) towards the end of the discharges.

4. Combination of RFs and balancing techniques

ML approaches have been widely used recently in the fusion context given their ability to learn complex patterns in problems where events evolve over multi spatio-temporal scales. They proved themselves useful tools in order to support first-principles physics models in the prediction of electron density and pressure profile shapes in tokamaks [37] and in disruption forecasting [19, 38, 39]. The latter are normally trained to predict with sufficient warning the likelihood that an instability will occur in the near future by treating the problem as a time-wise binary classification task. That is, individual time slices from a stable shot are labelled as negative (stable phase or far from instability), whereas an unstable discharge is split into two regions labelled as negative and positive (or close to instability), respectively. Although there is no ubiquitous way of defining the class label separation time in the unstable cases, it is quite common to choose 300 ms in advance of the instability, especially for disruption forecasting. In our case, such a choice would lead to an abundance of false positive warnings. In fact, previous analyses of the RKM revealed that while the collisionality drops in all the discharges due to increasing temperature, a pronounced turn towards higher $\langle \omega_E \rangle$ is generally observed at the end of the shot in the unstable cases [13]. Given the length of a typical NSTX discharge and that the RWM grows on a fast time scale (normally milliseconds), we have found that the optimal choice for the transition from negative to positive class is at $t^* = t_{\mathrm{RWM}} - t = 100$ ms.

One of the main consequences of this common setup is that the training data will be inevitably imbalanced with the risk of increasing the bias towards the majority class and jeopardizing the final performance; an effect that is further exacerbated by the fact that the minority class is often the one of interest. The simplest way of fixing this issue could be training the model via cost-sensitive learning (i.e. misclassified positive samples are heavily weighted), but we have found this approach leading to worse performance, as shown later. The use of random under/over-sampling [38] is an effective alternative to solely weighting the objective function but comes with some complications that must be taken into account. In fact, down-sampling by randomly deleting elements from the majority class might result in loss of valuable information, which means that its

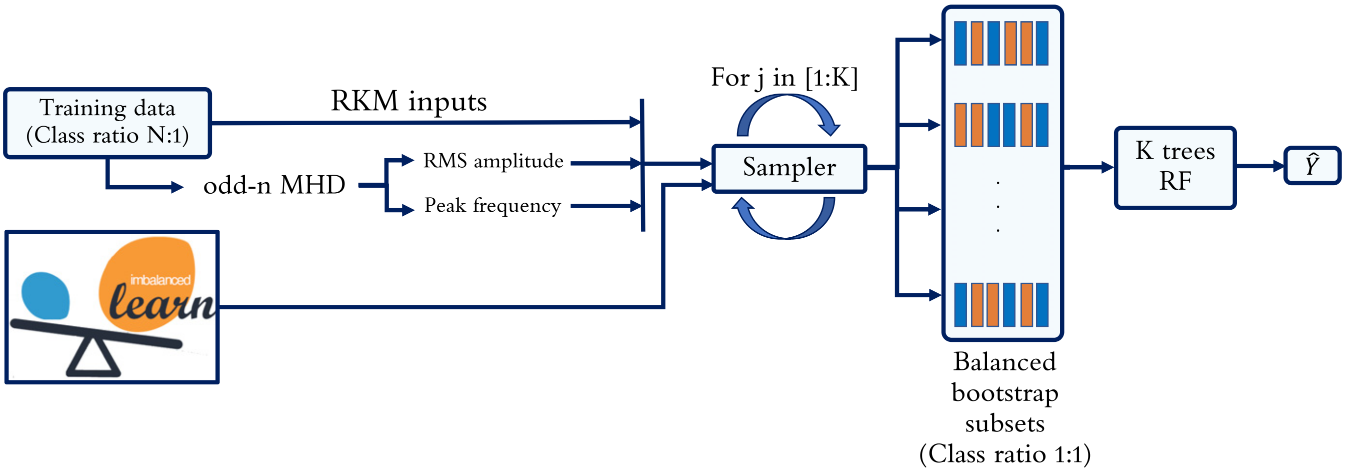

Figure 3. Architecture of the BRF-based RWM predictor's pipeline, in which the final output $\hat{Y}$ is the probability of being close to instability. The first five features in table 1 are directly fed into the model, whereas the odd-n MHD signal undergoes a preprocessing step to extract mode peak frequency and RMS amplitude. Each tree in the forest analyzes a different subset in which the training data has been balanced using a sampler chosen from the imbalanced-learn library. Note that the class ratio is roughly 16 : 1 for the best performing class label separation time of 100 ms.

Table 3. Performance of the proposed BUSRF classifier on individual time slices from the holdout test set compared to the classical WRF and other BRFs. The last column gives a sense of the time needed to train each model on an 8-core CPU for a budget of 1000 trials without early stopping.

Predictor	TPR	FPR	AUC	Cutoff	Training time (min)
WRF	76.8%	16.8%	0.895	0.410	156.5
BUSRF	92.4%	21.4%	0.918	0.450	61.3
BOSRF	87.0%	24.7%	0.889	0.275	178.4
SMOTERF	84.4%	22.2%	0.901	0.200	235.6
ADASYNRF	85.7%	22.9%	0.890	0.275	592.7

usage would be more suited for large datasets. On the other hand, up-sampling may increase the likelihood of overfitting since the model might learn rules that are tailored on replicas of the minority class.

Previous work [40] has focussed on the possibility of combining ensemble learning with random sampling in order to better learn from strongly imbalanced datasets. Ensemble learning is the combination of multiple machine learning methods to obtain a model that is more stable and less prone to overfitting. RFs ensure the required diversity and robustness and have been extensively used in several applications—including plasma physics and fusion energy—by proving to be robust tools for real-time disruption prediction on the DIII-D [23] and EAST [41] tokamaks, as well as in a cross-device fashion [20]. In the case of RFs, the imbalanced-learn package [42] provides an implementation of the balanced random forest (BRF) in which for each tree a bootstrapped sample from the minority class is drawn and then a random set of equal size is selected with replacement from the majority class. This process is different from classical sampling techniques because it involves reducing the number of samples at a tree-level rather than as a preprocessing step over the entire training set. Therefore, with a sufficient amount of estimators the ensemble will explore the dataset in its entirety regardless of the sampling strategy. Here we propose a customization of the original implementation of the BRF that allows plugging-in any sampling technique and choosing the one that performs best.

The learning phase has been performed via a stratified-by-shot cross-validator that not only splits the data in order to have non-overlapping sets of discharges in the different folds, but also takes into account the distribution of negative over positive time slices to stabilize the subsequent sampling step. Figure 3 illustrates the general pipeline followed during the training routine. The analysis carried out in the following section aims at showing the impact of several sampling techniques in terms of predictive performance as well as computational requirements. In particular, we have compared the classical RF using cost-sensitive learning (often referred to as weighted random forest, WRF [43]) with the ones that are regarded as being the most effective approaches, namely random under/over-sampling, SMOTE [44] and ADASYN [45].

5. BRFs performance

5.1. Results on individual time slices

The dataset has been split into training and testing sets comprised of 106 and 28 discharges, respectively. For each predictor, the best hyperparameters are chosen to be the ones that return the operating point on the cross-validation ROC curve which is closest to a perfect true positive rate (TPR = 1) and false positive rate (FPR = 0), with the hope that such parameters will perform well on the unseen test data. We have found the choice of this objective function to be the most reliable since its minimum represents the best compromise between a high TPR and a low FPR. In order to speed up the training process, we have employed a multivariate tree Parzen estimator (TPE) [46], which sequentially constructs a model to identify

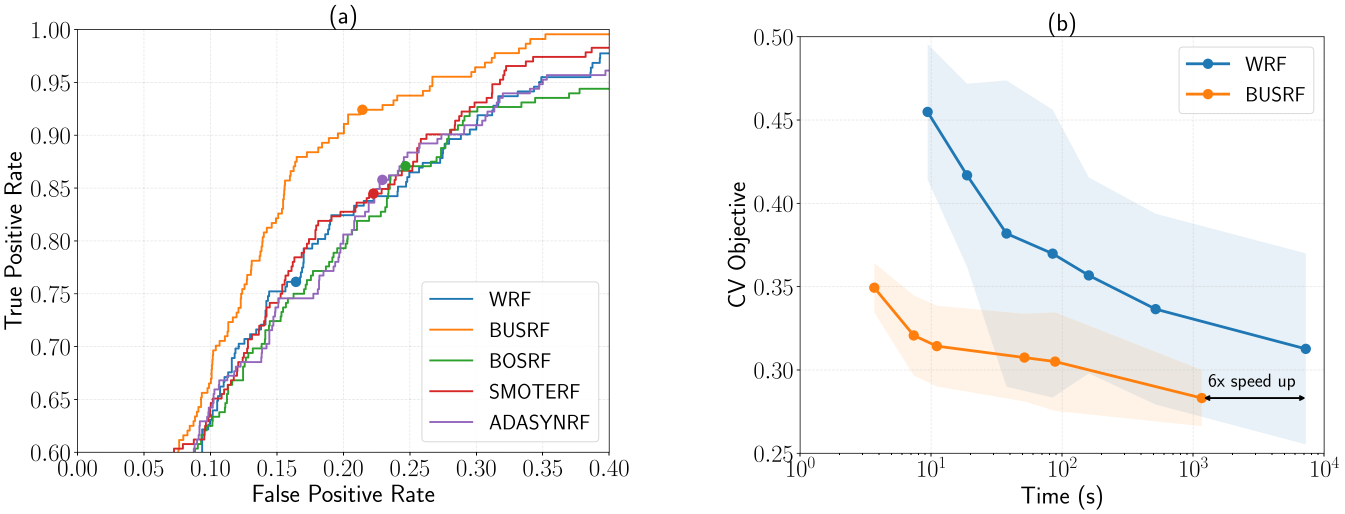

Figure 4. (*a*) Close-up look at the upper-left region of the test set ROC curves for the 5 trained classifiers. The optimal threshold is found via cross-validation and reported with a circle on the corresponding curve. (*b*) Learning curves for the WRF classifier and the proposed BUSRF approach. Although the absolute difference in the training time over 1000 trials is as reported in table 3, a post-hoc analysis has revealed that the objective converges to a minimum ∼6 times faster in the case of the BUSRF.

Table 4. Performance comparison of the tested predictors on a per-shot basis on the holdout test set, with the BUSRF highlighted in bold. Note that the BUSRF has the best detection capabilities without raising any early or late warnings.

Predictor	Unstable (/11)			Stable (/17)	Hyperparameters		
	Detected	Missed	Early	False positive	k_1	k_2	Δt_w (ms)
WRF	6	5	—	4	0.24	0.56	45
BUSRF	**10**	**1**	**—**	**2**	**0.36**	**0.71**	**60**
BOSRF	8	2	1	3	0.32	0.54	40
SMOTERF	8	3	—	2	0.36	0.42	35
ADASYNRF	7	3	1	2	0.26	0.58	60

promising new configurations based on the interdependencies between hyperparameters.

Table 3 summarizes the classification results obtained for prediction of stable or unstable individual time slices in the holdout test set using the different predictors. It is worth noticing that, although the WRF gives the lowest FPR, the BRF with random under-sampling (BUSRF) appears to have an edge over all the other approaches employed in terms of generalization performance and training time. As expected, the explored sampling techniques tend to improve the classification performance on the minority class, although the objective function gives the same weight to TPR and FPR. Fortunately, in all the cases the improvement in the TPR outweighs the worsening of the FPR, with a particularly favorable compromise for the BUSRF. This is an important aspect since in a fusion reactor we want no disruptions at the expense of a lower capacity factor. Although a false positive is undesirable, a wrongly triggered alarm would let active control steer the plasma back to a safe operational region.

In order to find the best operating point on the ROC curve, since the classifiers' predictions are continuous, we need to determine the best *cutoff* that maps the output to discrete values and allows for minimization of our chosen metric. Figure 4(*a*) displays a comparison of the ROC curves from the test set for all the trained classifiers.

On each curve, the optimal threshold chosen via cross-validation is indicated by a circle, which again shows that the BUSRF's estimate of the best threshold is the closest to the northwest corner. In this regard, it is also found that for the over-sampling techniques the optimal cutoff value is in the range 0.2–0.275, which tells us that these models are somewhat under-confident on their positive class predictions. This could be a consequence of the artificially generated data that might make these predictors more sensitive to boundary effects.

Finally, not only the searching method, but also the time required to train the underlying predictor strongly affects the optimization routine. A representative comparison between the learning curves for the WRF and BUSRF classifiers with a budget of 1000 trials is shown in figure 4(*b*). Rather than plotting the objective value for every single iteration, only those that have improved upon previous ones are marked. Evidently, the BUSRF converges to a minimum much faster than the WRF, with the latter also being generally more prone to noise [43], an effect that is visible by looking at the large standard deviation (blue shaded area) among the K folds. This

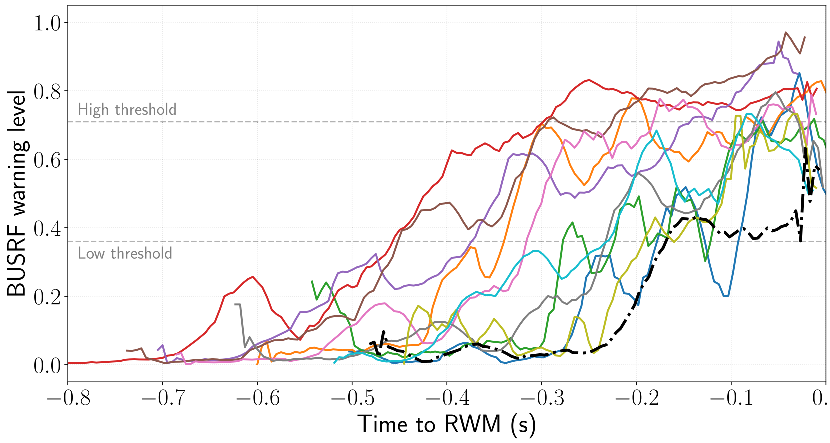

Figure 5. Predicted warning level for the 11 unstable discharges in the test set vs time to the human identified RWM instability. The superposed black dash-dotted line, although successfully identified by the RKM, is the only instability missed by the BUSRF in test phase.

issue is far less pronounced for the BUSRF even though the two predictors use exactly the same splitting. Although in our case the size of the dataset does not make the computational requirements of paramount importance, the potential advantage is the possibility of obtaining improved performance on large databases without having to deal with very long training times.

5.2. Results on a per-shot basis

The process of training the model on individual time slices would certainly give a sense of how suitable the method is to the problem at hand. However, while it is important to have a warning level for each time step in order to track the evolution of a discharge, the intuition is that a predictor should either trigger an alarm or not during the entire duration of a shot. The usage of a hard threshold has been observed to be highly sensitive to noise in the input signals as well as to random spikes in the predicted output. For this reason, previous works [38, 47] have proposed a 'softer' approach in which the model is re-trained with the addition of three new hyperparameters following the criterion that the alarm of impending instability is triggered only if the predictions stay above a low threshold (k_1) for a minimum amount of time (Δt_w) after having crossed a high threshold (k_2). An ideal warning criterion should counteract the issue of having alarms highly separated in time from the time of the actual instability as well as giving enough time for the plasma control system to act and avoid the instability. Although we could consider a warning raised any time during the shot as a success, we have decided to follow the DECAF RKM rule, in which if an alert is triggered more than 400 ms in advance without any related minor disruption, then it is considered too early. Moreover, in some of the tested predictors the alarm is raised too late (i.e. less than 10 ms before the RWM instability, regarded as missed).

Again, the algorithm that gave the lowest cross-validation objective was the BUSRF, with the model also resulting in the best test set performance by capturing 10 out 11 unstable RWMs (no late or early warnings) and just 2 out of 17 wrongly triggered stable discharges, as listed in table 4. We must point out that for the BUSRF one of the two false positives is a borderline case in which the alarm is triggered very early and then the warning level remains stable until the end of the shot, although it is unclear why this happened.

It is also interesting to notice that the best hyperparameters for the chosen predictor (BUSRF) slightly differ from the typical combinations in an RF-based disruption predictor, which further confirms that not only the machine but also the temporal evolution of plasma quantities associated with a specific instability affects the warning criterion. For example, in reference [47] the authors found that an ideal alarm should be triggered as soon as the output crosses the high threshold, with the low threshold at the lowest end of their search grid. On the other hand, Churchill *et al* [38] have partially confirmed this behaviour as their deep neural network triggers the alarm just 1 ms after exceeding a high threshold (0.96) which is quite large and also coincides with the low threshold. In our case the situation is somewhat in between, with k_1 being in the low-to-mid range, k_2 in the upper range, while Δt_w tells us that the predictor should wait at least 60 ms before triggering the alarm.

5.3. Comparison with the RKM and limitations

In addition to analyzing performance on the holdout test set, we are also interested in examining what we should expect from this new model and whether its predictions agree or not with the RKM. Although the warning times are slightly different, we have found the two models in general agreement apart from some peculiar cases. First of all, the RKM correctly identifies the unstable RWM missed by the BUSRF and predicts stability for one of the two discharges that are wrongly triggered by the ML model. On the other hand, the BUSRF appears to benefit from the usage of the absolute values of β_N, no-wall and with-wall limits rather than condensed into the C_β parameter, as well as from the addition of the odd-n MHD signal. In fact, in two unstable cases (curves in cyan and olive in figure 5) the BUSRF triggers an instability at the very end of the discharge with warning times of roughly 50 ms before the human identified RWM. Interestingly, both cases are not triggered by the RKM, possibly because of the low C_β or the high collisionality. In one unstable case (green curve) missed by the

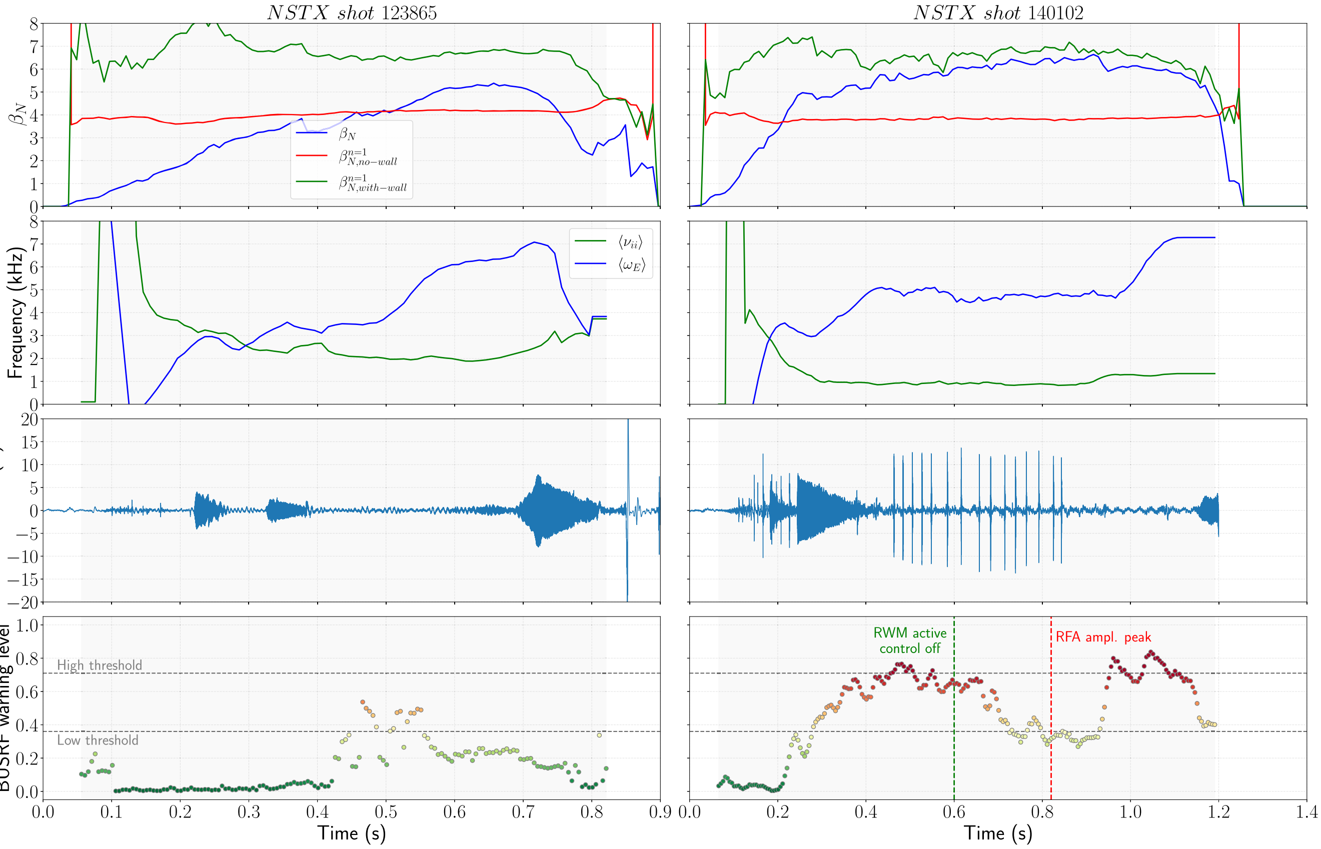

Figure 6. Comparison between a true negative discharge (Left) and a false positive one (Right). Both discharges are characterized by rising $\langle \omega_E \rangle$, decreasing $\langle \nu_{ii} \rangle$ and β_N crossing the no-wall limit. However, the very high β_N and low collisionality for shot 140102 seem to mislead our classifier, which wrongly triggers an early warning at $t = 0.441$ s.

RKM, the discharge evolved in the $\langle \nu_{ii} \rangle$ vs $\langle \omega_E \rangle$ space in what appears to be a stable range, but it is triggered 30 ms before the disruption by our model. Conversely, neither the BUSRF nor the RKM seem to identify stable discharges that evolve in what should be a highly unstable $\langle \nu_{ii} \rangle$, $\langle \omega_E \rangle$, β_N region.

The single test set shot belonging to this category is displayed on the right in figure 6 and compared to another stable discharge. By looking at the two plots, it is possible to highlight some of the limitations of the proposed model. The stable discharge on the left is characterized by steadily rising $\langle \omega_E \rangle$ along with β_N reaching roughly 5.2 at $t = 0.67$ s. No MHD activity is observed in the [0.4–0.7] s range, followed by increasing RMS amplitude and decreasing mode frequency that led to an LTM during the plasma current rampdown at $t = 0.851$ s. In contrast, shot 140102 had a broader rotation profile, ω_Φ, and remained RWM stable as ω_Φ was reduced by $n = 3$ magnetic braking [48]. For this specific shot, the proximity to marginal stability as rotation is reduced was evaluated by active MHD spectroscopy. Active $n = 1$ RWM feedback control was turned off at around 0.6 s, followed by the application of an $n = 1$, 40 Hz co-NBI propagating tracer field to evaluate the resonant field amplification (RFA) due to the high β_N level. Although during this time the plasma remained at a high and relatively constant β_N of roughly 6.2, the RFA amplitude peaks at around 0.82 s and then decreases as ω_Φ is reduced. This denotes that the plasma first approaches, then departs from RWM marginal stability [49]. Evidently, these early signs of possible RWM instability led the BUSRF to trigger an early false alarm despite the presence of rotating MHD activity at the beginning of the shot. Interestingly, the BUSRF is then picking up in advance that something is changing in the plasma evolution, with the warning level monotonically decreasing as the plasma rotation is reduced. Unfortunately, the model raises another warning right at the end of the shot, corresponding to a steep rise in $\langle \omega_E \rangle$. This analysis further shows that ML models should be used in conjunction with plasma diagnostics and existing control systems to define paths for improvement, especially when it comes to marginal cases such as the one presented here. Although just one of these cases is observed in the testing set, 4 other similar discharges are also wrongly triggered during the cross-validation process. Presumably, both models are also not capturing other stabilizing aspects such as the effect of energetic particles (i.e. temperature and pressure of the thermal ions). One possibility in the future could be using fast ion pressure profiles produced by the NUBEAM module of the TRANSP code [50], or reconstructed by a neural network [51], and then injecting this additional data into a larger RWM stability module.

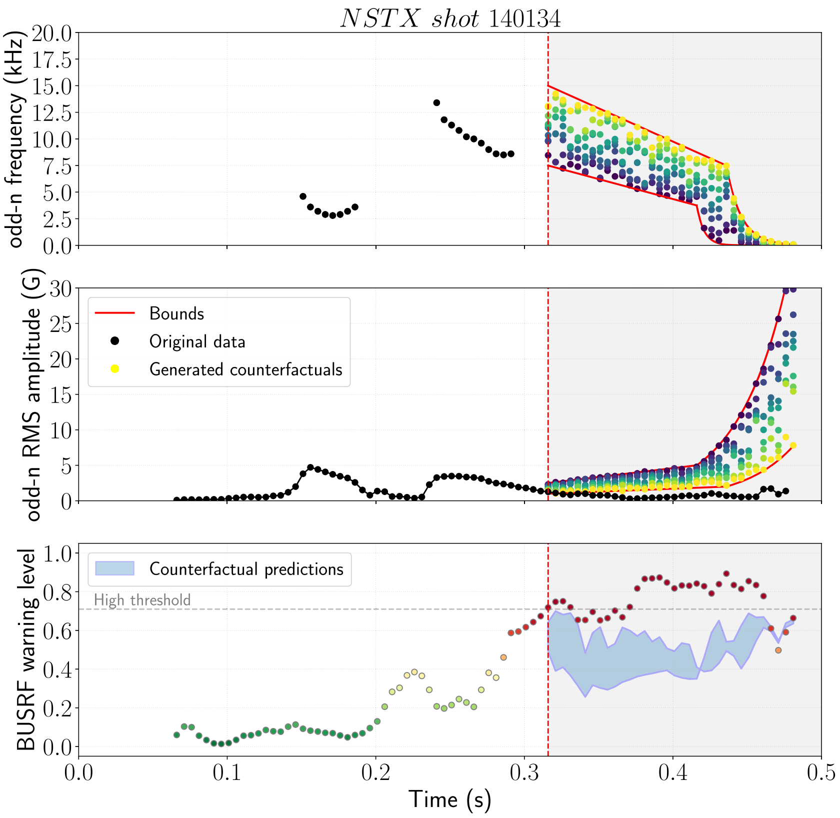

Figure 7. Generation of odd-n MHD counterfactuals for NSTX shot 140134, which was experimentally RWM unstable. In the top and middle panels are displayed the original and simulated odd-n activities in black and colour, respectively. Each colour represents one of the ten possible levels of MHD activity. The bottom panel shows the original BUSRF predictions as coloured dots, while the blue shaded area represents the range of counterfactual predictions if strong MHD activity was present. Note that, for the sake of visualization, only the frequencies where the RMS amplitude is at least 2 G are displayed in black in the top panel.

Table 5. Equations for the lower and upper bounds used to constrain the DiCE algorithm for NSTX shot 140134. The exponential starting points are separated by 20 ms in order to avoid convergence of the bounds and the eventual breaking of the required diversity among counterfactuals.

Range	Frequency (kHz)		RMS amplitude (G)	
	Lower	Upper	Lower	Upper
$t_0 \leqslant t < t_1$	$7.5 - 3.75\,\frac{t-t_0}{t_1-t_0}$	—	—	$2.5\left(1+\frac{t-t_0}{t_1-t_0}\right)$
$t_0 \leqslant t < t_2$	—	$15 - 7.5\,\frac{t-t_0}{t_2-t_0}$	$1+\frac{t-t_0}{t_2-t_0}$	—
$t \geqslant t_1$	$3.75e^{-200\,(t-t_1)}$	—	—	$5e^{30\,(t-t_1)}$
$t \geqslant t_2$	—	$7.5e^{-100\,(t-t_2)}$	$2e^{30\,(t-t_2)}$	—

6. Interpretation of the warning level via diverse counterfactuals

Post-hoc interpretation of complex models are key for physicists to understand algorithmic predictions and judge whether they agree with the underlying physics or not. In the RKM, it is possible to follow plasma trajectories over time in the $\langle \nu_{ii} \rangle$, $\langle \omega_E \rangle$ space by drawing contours of $\gamma\tau_w$ [13]. In our case, the model's decision boundary lies on a sevev-dimensional manifold, making us unable to visualize it unless we performed a scan of two features at a time while keeping the others at a constant or highly representative value. Other works [23, 47] have explained the decisions made by multi-dimensional RF classifiers by decomposing the predictions into feature contributions. Here we propose the use of an alternative and model-agnostic method of interpreting the outcome based on counterfactuals, or *what-if* explanations. The concept of counterfactuals relies on the generation of hypothetical combinations of the input features that would return a specified output. In our case, the most obvious choice would be the generation of plasma parameters that produce predictions on the stable side of the decision boundary. Such a framework should ideally respect three conditions, which are: the ability to suggest a *diverse* set of combinations, the possibility of practically performing these changes (i.e. *proximity* to the original input space), and the consideration of the causal links in a real case scenario.

The diverse counterfactual explanations (DiCE) algorithm has proved to satisfy all three requirements and provides the right flexibility to take into account user-defined constraints and experts' knowledge. For the sake of brevity, we will not report its complete formulation [29],[3] but rather highlight the main points and show how to customize it to the problem at hand. For non-gradient based ML methods, DiCE uses genetic programing to minimize an objective function over the entire set of potential candidates as follows:

$$\mathcal{C}_1(\mathbf{x}) = \underset{c_1,\ldots,c_k}{\arg\min}\left[\underbrace{\frac{1}{k}\sum_{i=1}^{k}\mathcal{L}(f(\mathbf{c_i}),y)}_{\text{Hinge loss}} + \underbrace{\frac{\lambda_1}{k}\sum_{i=1}^{k}\mathcal{D}_1(\mathbf{c_i},\mathbf{x})}_{\text{Proximity}} - \underbrace{\lambda_2\mathcal{D}_2(\mathbf{c}_1,\ldots,\mathbf{c_k})}_{\text{Diversity}}\right], \tag{2}$$

where $c_1,\ldots,c_k$ is a set of k counterfactual examples, $\mathcal{L}$ is the hinge loss between the desired output y and the ML model's predictions ($f(\mathbf{c_i})$) associated with the set of proposed counterfactuals, $\mathcal{D}_1$ is a measure of perturbation defined as the normalized average feature-wise l_1-distances from the original input ($\mathbf{x}$), $\mathcal{D}_2$ is a diversity metric based on DPP (see appendix) that avoids repetition of the same $\mathbf{c_i}$ in a set of given size k, and λ_1 and λ_2 can be considered as regularization coefficients that are used to balance the impact of diversity and proximity. The hinge loss guarantees a smaller perturbation of the original set of features $\mathbf{x}$ since in a single-threshold model a counterfactual only needs to be on the other side of the decision boundary in order to be valid.

As a starting point, we have customized the DiCE algorithm in several ways. First, physics-informed counterfactuals are generated in order to gain understanding into the model's decisions. For example, based on physics knowledge we expect that the usage of the odd-n MHD signal should improve the model's capabilities, but we do not know *a priori how* it is actually helping. Second, this approach can be modified and potentially used in a real-time control algorithm by stepping down the most influential parameters, such as β_N, in order to see which levels keep the plasma RWM stable. Additionally, in order to speed up the counterfactual generation process, a TPE-based optimizer was added to the DiCE framework.

6.1. Usage of counterfactuals to generate hypothetical odd-n MHD activity

Altogether, the extracted mode frequency and RMS amplitude cover roughly 20% of the relative feature importance, hence we can assume that these features definitely influence the decisions made by our classifier, as expected. Notwithstanding the fact that only some of the RWM-stable discharges had strong MHD activity, we can still explore the ability of the BUSRF in understanding how RWM stability is affected by the presence of low order MHD modes, in particular slowing ones. As shown previously in figure 2(*b*), the MHD mode spectrum from a toroidal array of magnetic pickup loops for an unstable RWM shows no mode activity, indicative of stable rotating MHD. As an additional example, figure 7 displays one of the analysed RWM unstable shots, correctly identified by the BUSRF at $t_{\text{trigger}} = 0.376$ s. By looking at the black points in the top and middle panels, one can notice that the absence of MHD activity towards the end of the shot is followed by the warning level (bottom panel) rising above the high threshold for the first time at $t_0 = 0.316$ s, indicated by a change to a grey background.

Starting from this point, we have asked the algorithm to produce a set of 10 possible combinations of MHD frequencies and amplitudes per time step that would have prevented the BUSRF from triggering an RWM warning. Since the original main loss only ensures that a counterfactual lies below or above a fixed threshold of 0.5, we chose to replace the hinge loss in equation (2) with a custom step function that penalizes counterfactual predictions above the high threshold, as follows:

$$\mathcal{L}_{\mathcal{H}}(f(\mathbf{c})) = \frac{1}{k}\sum_{i=1}^{k}\mathcal{H}(f(\mathbf{c_i})) \quad \text{where}$$
$$\mathcal{H}(f(\mathbf{c_i})) = \begin{cases} 0 & \text{if } 0 \leqslant f(\mathbf{c_i}) < k_2 \\ \lambda_0 & \text{if } f(\mathbf{c_i}) \geqslant k_2 \end{cases} \tag{3}$$

hence the optimization will tend to prioritize DPP-diversity while keeping counterfactuals below the high threshold.

Previous DECAF characterization of LTMs in NSTX discharges has found that before the disruption the frequency decreases and linearly halves, then it bifurcates and abruptly drops towards zero [14]. An opposite trend is generally observed with the RMS amplitude. In order to make these hypothetical situations as realistic as possible, in this application the algorithm has been constrained within the red curves, which follow these considerations and whose expressions can be found in table 5. Specifically, for this shot the bounds change from linear trends to exponential decays (or growths) at $t_1 = 0.416$ s and $t_2 = 0.436$ s, respectively. Moreover, we have found that the choice of $\lambda_0 = 1$ and $\lambda_2 = 4$ was the best in order to generate a diverse set of opposite MHD scenarios while keeping counterfactuals inside the bounds. For this specific application, proximity is not a requirement, hence λ_1 was set to 0.

The generated MHD activity in the top and middle panels of figure 7, when applied to the RWM unstable discharge, produces the counterfactual BUSRF warning level shown in the bottom panel, which is now stable. The choice of setting the initial bounds to 7.5 and 15 kHz is just an example of such application. In fact, we have also tested the same approach on other 5 unstable discharges by narrowing or widening the initial gap between 5 and 20 kHz, which are values that come from experimental observation of MHD activity in NSTX discharges [14]. On the other hand, the RMS amplitude bounds

[3] An easy-to-use implementation can be found at https://github.com/microsoft/DiCE

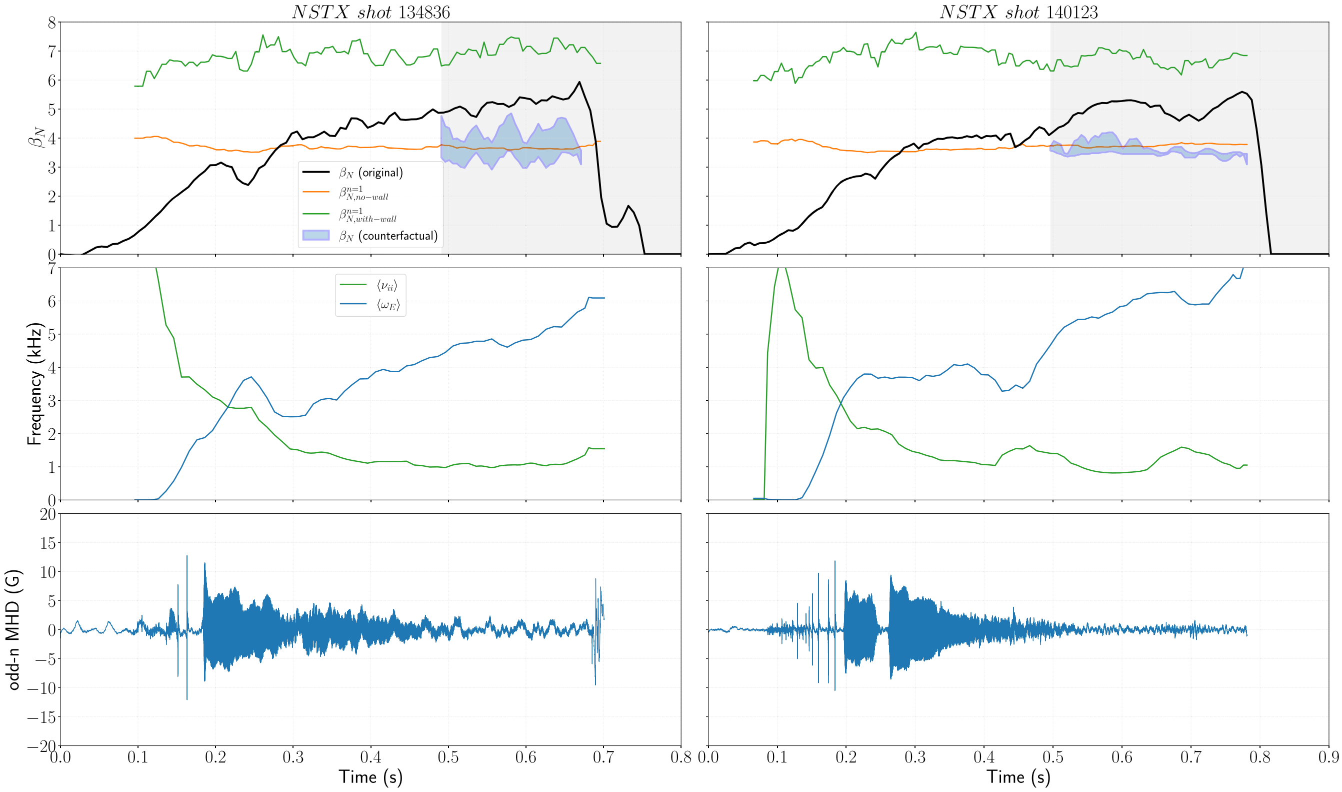

Figure 8. Safe normalized beta regions suggested by the algorithm for shots 134836 (Left) and 140123 (Right). The background changes to grey as soon as the BUSRF warning level rises above the low threshold, corresponding to the time at which the algorithm starts generating the blue shaded counterfactuals.

have been left unchanged in all the combinations as well as the halving times, which are set to 100 and 120 ms, respectively. In all the experiments DiCE was able to find a set of counterfactuals that would have kept the warning level below k_2 until the end of the discharge. Therefore, the counterfactual approach is confirming the physics understanding that the global RWM mode cannot grow to instability while local MHD activity is present, growing in amplitude and slowing to potentially disrupt the plasma itself. The same technique might be used to simulate MHD scenarios in which both the bounds and the time scales can vary based on NSTX experience as well as to evaluate the reliability of our model if run in parallel with the DECAF LTM warning module. It must be stated that such alternative scenarios might still lead to a disruption, which however should not be caused by an unstable RWM and we expect our model to not trigger any alarm. However, other possible causes of disruptions forecast by the DECAF code could be evaluated separately given the parameter space spanned by the counterfactual analysis for RWM.

6.2. ML-informed safe scenarios for potential real-time control

The second application of this approach lays the basis for a potential usage of counterfactuals for real-time control of the main RWM instability drivers, which are β_N and rotation. Normalized beta control has already been accomplished in NSTX [52] by means of a proportional-integral-derivative gain controller that changes the injected power in order to achieve a user-defined β_N level. In NSTX, all the beams are pointed in the same direction, hence in addition to more heating, more beam power necessarily means more plasma rotation. Moreover, because of the different angles at which the different beam sources were injected, an increase or decrease in power in one beam or the other would also influence pressure peaking and internal inductance (the two are correlated). For example, the beam that penetrated farther into the plasma core would tend to increase pressure peaking whereas one that was more towards the edge would tend to decrease it. Therefore, such a modulation would also modify the no-wall and with-wall beta limits, which would in turn play a role in the global RWM stability. For example, a less peaked pressure profile generally increases the with-wall limit allowing access to higher β_N. Although other parameters could change simultaneously as β_N is stepped down, an ML-informed safe β_N level represents a valuable piece of information that could be input to a more sophisticated control method [53] already tested and planned for future NSTX-U operation. Alternatively, an all-encompassing counterfactual generation algorithm for RWM stability control can be built, but this is mostly future work.

For the present purpose, we will generate the time-wise β_N levels that would have kept the plasma RWM stable for a set of NSTX discharges for which the BUSRF had triggered an alarm, and we will do so by replacing equation (2) with an objective function that constrains counterfactuals in the probability domain rather than in the input space. Since a high β_N is desirable to achieve good fusion performance, we will find

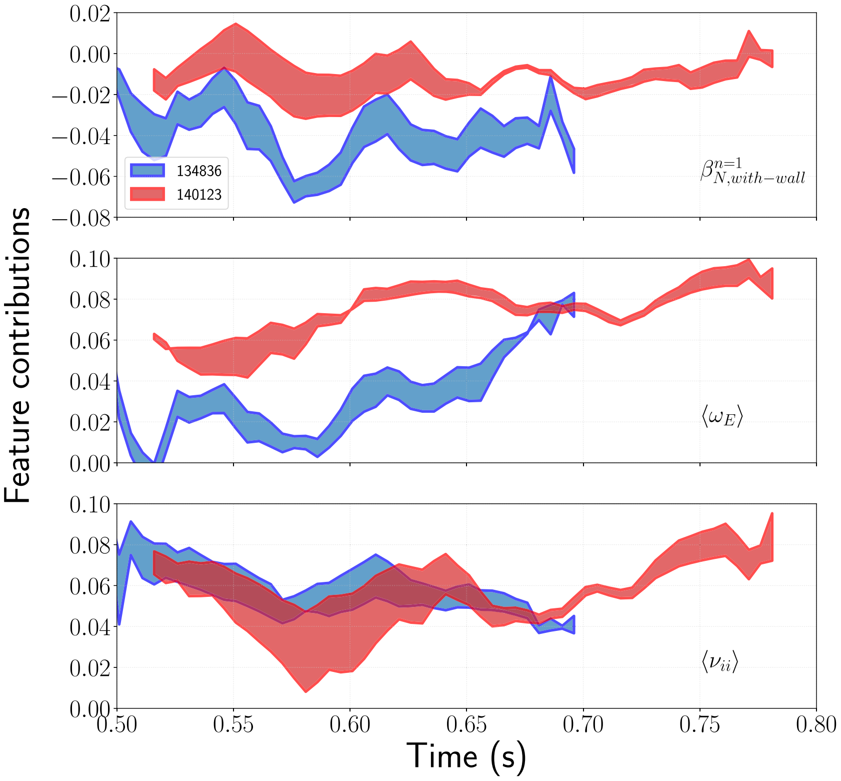

Figure 9. The region in which DiCE operates (grey in figure 8) is analysed in terms of the most relevant feature contributions while β_N is reduced. Each error bar represents the contribution associated with the two levels β_N^0, β_N^1 found by minimizing equation (4). The collisionality dropping below 1 kHz for shot 140123 allows access to higher rotation, but also results in a heavier impact of the latter towards the final prediction.

the β_N region that keeps the warning level between 0.5 and the high threshold, k_2, as follows:

$$\mathcal{C}_2(\beta_N^0, \beta_N^1) = \underset{(\beta_N^0,\beta_N^1)}{\arg\min}\left[|f(\beta_N^0) - 0.5| + |f(\beta_N^1) - k_2|\right] \text{ with } 0.5 \leqslant f(\beta_N^i) < k_2. \quad (4)$$

Apart from speeding up the counterfactual generation process, this objective function also maintains the diversity and proximity properties. In fact, given the desired probability range we expect the two β_N values to not coincide as well as to lie in the high beta region.

Figure 8 shows the application of such a method to two similar high beta discharges in which the RWM became unstable in the chain of events leading to a disruption. The main characteristic that one can notice is that the blue shaded safe β_N region lies around/slightly above the NN-modeled no-wall limit. The behaviour is not surprising, as the goal was to find a safe operating region and the counterfactual approach is confirming that, to first order, increasing beta above the no-wall limit increases the disruption probability. Two other things emerge from the analysis of the two discharges as well. First of all, the slightly higher with-wall limit for shot 134836—indicative of a less peaked pressure profile—seems to partially justify the maximum safe β_N of roughly 4.8, somewhat above the no-wall limit, in the grey counterfactual region. On the same grounds, one can notice that the blue area moves upwards between 0.55–0.6 s for shot 140123, corresponding to the times at which the with-wall limit is almost 7. Second, while both discharges are characterized by decreasing collisionality, the rise in $\langle\omega_E\rangle$ is more pronounced for shot 140123, in which it crosses 6 kHz at around 0.62 s. This may further explain the much narrower safe β_N operating space towards the end of this discharge.

All these considerations can be confirmed by producing the time-wise feature contributions [23] for the BUSRF classifier. As a reminder, this analysis essentially decomposes the prediction into a bias term plus the sum of each feature's contribution[4]. The bias term in this case is 0.5 since trees in the BUSRF are trained on a perfectly balanced dataset. We have found that the contributions of both $\beta^{n=1}_{N,\text{no-wall}}$ and the MHD activity are roughly the same between these two specific discharges since the evolution in time of these two features is almost identical between the two shots. Therefore, the corresponding contributions do not inform us about why the counterfactual β_N region is different between the two discharges. Conversely, we compare the effect of $\beta^{n=1}_{N,\text{with-wall}}$, $\langle\omega_E\rangle$ and $\langle\nu_{ii}\rangle$ towards the warning level in the following figure 9.

Since the counterfactual generated β_N is a region that changes in time, the individual contributions are generated as error bars as well. By looking at the top panel, one can notice

[4] For a dataset of N features the warning level breakdown would be prediction $=$ bias $+ \sum_{i=1}^{N}$contrib$_i$

that the higher with-wall limit for shot 134836 is effectively damping the warning level, in particular around 0.57 s when the counterfactual β_N reaches its maximum. This effect is less pronounced in the red shaded area, related to a $\beta_{N,\text{with-wall}}^{n=1}$ that barely reaches 7. Moreover, it is worth noticing the contributions of $\langle \nu_{ii} \rangle$ and $\langle \omega_E \rangle$. Physics expectation based on MISK calculations has shown that RWM instability is possible at very low rotation and also at a higher rotation level between stabilizing resonances. This higher rotation level at which the RWM becomes unstable tends to increase as collisionality decreases [49]. This appears to be consistent with the behaviour of shot 140123. As $\langle \nu_{ii} \rangle$ drops below 1 kHz access to higher rotation is allowed, so $\langle \omega_E \rangle$ steeply rises. Meanwhile collisionality starts to oscillate. One can see that the contribution of $\langle \nu_{ii} \rangle$ first decreases and then increases again, which confirms that it is not possible to assume that stability decreases monotonically with collisionality [7]. The joint effect of these three contributions is an increase in the warning level that, in the counterfactual scenario, can only be damped by a reduction in the β_N level below the no-wall limit, as seen on the top right panel in figure 8. Overall, the warning level is 0.05 to 0.15 higher for this discharge, hence the lower and narrower counterfactual region suggested by the DiCE algorithm.

As was previously mentioned, β_N modulation also influences rotation in a device like NSTX. Therefore, in the future this approach can be applied in a way that takes into account, for example, the simultaneous variation of β_N and $\langle \omega_E \rangle$. In addition, the reliability of any counterfactual scenario relies on the accuracy of the underlying model. Therefore, it is really important building a robust counterfactual generator as well as perfecting the RWM stability forecaster. The implementation of ML-based controllers has been gaining attention in recent years with through the usage of reinforcement learning, for example to determine the optimal control scenario to obtain a constant user-defined β_N level in the KSTAR tokamak [54]. The direction we have taken in the present work follows the observation that the disruptive β_N is not constant when it comes to RWM stability and the usage of counterfactuals might lay the basis for a control algorithm that takes this variability into account. Moreover, recent research [55] has shown that contribution-based explanations can be jointly used with counterfactual generation algorithms and disclose which features are necessary or sufficient towards the model's decisions.

7. Discussion and conclusions

The prediction of impending instabilities is crucial for the safe operation of future commercial fusion devices. The combination of physics knowledge and the capabilities of ML algorithms is arguably one of the best solutions to this grand challenge. Recently, particular attention has been given to interpretable ML models, such as RFs, which can easily give an insight into the underlying causes of instabilities by providing the importance of each input feature as well as their contribution to the final output. A new RWM stability forecaster based on the RF approach has been developed here to predict whether the RWM will go unstable or not for a set of human-labelled NSTX discharges. The present approach outperforms classical weighted RFs by including a random per-tree undersampler to balance individual time slices. The algorithm is a valuable addition to DECAF and can be regarded as a supporting tool for the existing RKM model, since a processed signal indicative of the presence of low frequency MHD activity has been added to the input features.

Additionally, a model-agnostic tool based on counterfactuals has been tested for the first time to explain the effect of the hypothetical presence of MHD activity as well as to simulate β_N levels that would have kept the plasma RWM stable. Regarding the first usage, it has been observed that the underlying model is understanding that strong MHD activity is generally unobserved when the RWM goes unstable, as expected by prior experimental evidence. On the other, suggested counterfactuals have interestingly shown that in discharges where the RWM warning level rises above 70%, the first line of defense is to reduce β_N in a range that is slightly above the no-wall limit.

Both the BRF and the counterfactual approaches are applicable to other tokamaks as well. While it may be necessary to retrain the underlying ML model on data from each machine, it should be noted that the PGNN-determined $\beta_{N,\text{no-wall}}^{n=1}$ from NSTX data was applied with some initial success to the MAST tokamak [27]. Although more research is needed to determine the reliability of cross-device applications, the physics in the kinetic stability terms, $\langle \omega_E \rangle$ and $\langle \nu_{ii} \rangle$, are applicable to other devices as well [12]. For example, comparisons of the theory to dedicated DIII-D experiments were successful. Other terms, like the no-wall and with-wall limits, are based on physics parameters, such as the aspect ratio, that can scale to other machines. Finally, recent calculations [56] for the MAST-U tokamak show potential high beta stability regions that could provide valuable information to better characterize RWM stability across multiple devices.

There are various elements in this approach that may be improved. First of all, the underlying RF model might be refined with the inclusion of the effect of energetic particles from TRANSP NUBEAM runs, which would likely help in reducing the FPR. Secondly, although the current usage of counterfactuals is encouraging, it still needs to include interdependencies between the main plasma parameters. Work in the near future will cover the implementation of counterfactual explanations to further gain understanding of RWM stability, exploring actionable scenarios and implementing a more sophisticated β_N/rotation ML-based controller.

Acknowledgments

The authors acknowledge the help of J D Riquezes for the essential knowledge provided on rotating MHD modes in NSTX, and J J Yang for the useful comments on the importance of improving the model's success rate. This research was supported by EPSRC Grants EP/R025290/1 and EP/P02243X/1 (University College London), and by US Department of Energy under Contract DE-SC0018623 (Columbia University). The digital data for this paper can be found at http://arks.princeton.edu/ark:/88435/dsp015q47rr88m

Appendix A

A.1. Details about the reduced physics models and the input features to the BRF classifiers

A.1.1. PGNN-informed no-wall and heuristic with-wall limits. The equation for the no-wall beta limit used in this work (see table 1) is from reference [27]:

$$\beta_{N,\text{no}-\text{wall}}^{n=1} = \sqrt[3]{\left[\left(\sum_{x_i} a_{\beta_N,x_i} w_{\beta_N,x_i}\right) - a_0\right]\left[\sum_{x_i} \frac{a_{\beta_N,x_i} w_{\beta_N,x_i}}{\beta_{N,\text{bnd}}(x_i)^3}\right]^{-1}}, \tag{5}$$

where a_{β_N,x_i}, w_{β_N,x_i} and a_0 are optimized coefficients that can be found in table 6 of reference [27]. Whereas the $\beta_{N,\text{bnd}}(x_i)$ terms are the three decision boundaries found by the PGNN trained on DCON calculations as follows:

$$\beta_{N,\text{bnd}}(l_i) = 4.91 \exp\left(-\left(\frac{l_i - 1.17}{1.14}\right)^2\right) + 0.21 \exp\left(-\left(\frac{l_i - 0.27}{0.39}\right)^2\right) \tag{6}$$

$$\beta_{N,\text{bnd}}(A) = -4.14A^3 + 13.47A^2 - 14.95A + 10 \tag{7}$$

$$\beta_{N,\text{bnd}}(F_p) = 2.56 \exp\left(-\left(\frac{F_p - 4.06}{2.59}\right)^2\right) + 3.10 \exp\left(-\left(\frac{F_p - 0.43}{4.29}\right)^2\right), \tag{8}$$

where l_i is the internal inductance, A is the aspect ratio, and $F_p = p_0/\langle p\rangle$ is the pressure peaking factor. For the with-wall limit:

$$\beta_{N,\text{with}-\text{wall}}^{n=1} = 0.75 + 12.5/F_p \tag{9}$$

which comes from a fit to the DCON projected with-wall limit found for NSTX-U [57] and adapted to the operational limit seen in NSTX [13].

A.1.2. The kinetic term δW_k. The full equation and the reduced model for δW_K from reference [13] are reproduced here, for reference, although they are not used in the present work. Rather the important parameters $\langle\omega_E\rangle$ and $\langle\nu_{ii}\rangle$ were identified and included as inputs to the ML algorithm (see table 1):

$$\delta W_k \sim \int_0^\infty \left[\frac{\omega_{*N} + (\hat{\varepsilon} - 3/2)\omega_{*T} + \omega_E - \omega_r - i\gamma}{\bar{\omega_D}\hat{\varepsilon} + l\bar{\omega_b}\hat{\varepsilon}^{1/2} - i\bar{\nu}\hat{\varepsilon}^{-3/2} + \omega_E - \omega_r - i\gamma}\right]\hat{\varepsilon}^{5/2}e^{-\hat{\varepsilon}}\,\mathrm{d}\hat{\varepsilon}, \tag{10}$$

where $\hat{\varepsilon}$ is the particle energy normalized by temperature, ω_{*N} and ω_{*T} are the density and temperature gradient components of the ion diamagnetic drift frequency, ω_E is the $E \times B$ frequency, ω_D the precession drift frequency, l the bounce harmonic, ω_b the bounce frequency, and ω_r and γ the real frequency and growth rate of the mode. This complex form was reduced, through various assumptions, to functional expressions for the real and imaginary, precession and bounce δW_k terms, as follows:

$$\delta W_k = a\frac{\langle\omega_E\rangle}{\omega}\exp\left[-\frac{(\langle\omega_E\rangle/\omega - b)^2}{2c^2}\right], \tag{11}$$

where a, b and c are functions of $\langle\nu_{ii}\rangle/\omega$, with ω being the normalizing frequency $\omega_D = 2$ kHz or $\omega_b = 10$ kHz for each respective piece of δW_k, and $\langle\rangle$ represents an average value inside the profile pedestal.

In the above equations as well as in the present work, the $E \times B$ profile is related to plasma rotation through $\omega_E = \omega_\Phi - \omega_{*N} - \omega_{*T}$ [49], with the diamagnetic drift frequency terms obtained as follows:

$$\omega_{*N} = -\frac{T_i}{en_i}\frac{\mathrm{d}n_i}{\mathrm{d}\Psi} \tag{12}$$

$$\omega_{*T} = -\frac{1}{e}\frac{\mathrm{d}T_i}{\mathrm{d}\Psi}, \tag{13}$$

where the above gradients are taken with respect to the equilibrium poloidal magnetic flux, Ψ, in units of Wb.

A.2. Definition of the diversity term for the counterfactual generator

The counterfactual generation process is a compromise between actionability, proximity and diversity. As mentioned previously, proximity is not a necessary requirement in the case of alternative MHD activity generation, hence we will focus here on showing how diversity has been derived. In reference [29], counterfactual diversity is achieved by computing the DPP of the pairwise inverse distance matrix, $\mathcal{M}$, between the proposed counterfactuals. In our case, we have found that the exponential of the distance works best. Hence each term of $\mathcal{M}$ is obtained as follows:

$$\mathcal{M}(i,j) = \exp\left(-dist(\mathbf{c_i}, \mathbf{c_j})\right) \quad \text{with} \quad dist(\mathbf{c_i}, \mathbf{c_j}) = \frac{1}{d}\sum_{p=1}^{d}\frac{|\mathbf{c_i} - \mathbf{c_j}|}{MAD_p}, \tag{14}$$

where d is the number of continuous features (7 in the present work) and *MAD* is the feature-wise median absolute deviation to take into account that features span over different ranges. The DPP provides a fast and efficient way to capture negative correlation with respect to a similarity measure, which can in turn be used to quantify the diversity within a set of feature instances. It is clear that the more diverse two counterfactuals are, the smaller the corresponding inverse exponential distance would be. As an example, we have reported in table 6 the output of the algorithm after minimizing equation (2) for shot 140134 (see figure 7) at $t = 0.366$ s. In the top row, the original features are highlighted in bold. In this case, $\mathcal{M}$ was a 10×10 matrix in which the non-diagonal elements were essentially determined by the features that we chose to change, corresponding to the last two columns. Conversely, since dist $(\mathbf{c_i}, \mathbf{c_j}) = 0$ for $i = j$, the entries on the leading diagonal are all ones. The first three columns of the kernel $\mathcal{M}$ associated to the

Table 6. Example of the counterfactuals generated by DiCE for shot 140134 at $t = 0.366$ s. The last two columns are the suggested set of alternative scenarios that would have prevented the warning level from rising above 0.7 and triggering an RWM alarm.

		Unchanged features					Features to change	
	t	β_N	$\beta^{n=1}_{N,\text{no-wall}}$	$\beta^{n=1}_{\text{with,no-wall}}$	$\langle\omega_E\rangle$	$\langle\nu_{ii}\rangle$	Odd-n MHD freq.	Odd-n MHD RMS
	0.366 s	**4.29**	**4.03**	**5.54**	**4.05**	**2.89**	**0.20**	**0.46**
Proposed counterfactuals		—	—	—	—	—	6.54	3.50
		—	—	—	—	—	6.74	3.34
		—	—	—	—	—	6.88	3.24
		—	—	—	—	—	7.62	3.03
		—	—	—	—	—	7.78	2.85
		—	—	—	—	—	7.84	2.78
		—	—	—	—	—	8.88	2.46
		—	—	—	—	—	11.42	2.25
		—	—	—	—	—	11.62	2.18
		—	—	—	—	—	11.80	1.41

counterfactuals in table 6 are reported below:

$$\mathcal{M}_{(t=0.366\text{ s})} = \begin{bmatrix} 1.0 & 2.5\cdot10^{-1} & 1.0\cdot10^{-1} & \dots & \dots \\ 2.5\cdot10^{-1} & 1.0 & 3.9\cdot10^{-1} & \dots & \dots \\ 1.0\cdot10^{-1} & 3.9\cdot10^{-1} & 1.0 & \dots & \dots \\ 1.5\cdot10^{-3} & 5.8\cdot10^{-3} & 1.5\cdot10^{-2} & \dots & \dots \\ 4.3\cdot10^{-4} & 1.7\cdot10^{-3} & 4.3\cdot10^{-3} & \dots & \dots \\ 2.8\cdot10^{-4} & 1.1\cdot10^{-3} & 2.8\cdot10^{-3} & \dots & \dots \\ 7.0\cdot10^{-7} & 2.8\cdot10^{-6} & 7.0\cdot10^{-6} & \dots & \dots \\ 1.3\cdot10^{-12} & 5.1\cdot10^{-12} & 1.3\cdot10^{-11} & \dots & \dots \\ 4.0\cdot10^{-13} & 1.6\cdot10^{-12} & 4.0\cdot10^{-12} & \dots & \dots \\ 2.6\cdot10^{-14} & 1.0\cdot10^{-13} & 2.6\cdot10^{-13} & \dots & \dots \end{bmatrix} \quad (15)$$

in which one can notice that the per-column similarity metric decreases the farther we move from the $i = j$ entry, indicative of increasing diversity.

ORCID iDs

A. Piccione https://orcid.org/0000-0002-2746-0723

References

[1] Bondeson A. and Ward D.J. 1994 *Phys. Rev. Lett.* **72** 2709–12

[2] Chu M.S. and Okabayashi M. 2010 *Plasma Phys. Control. Fusion* **52** 123001

[3] Sabbagh S.A. *et al* 2002 *Phys. Plasmas* **9** 2085–92

[4] Berkery J.W., Sabbagh S.A., Reimerdes H., Betti R., Hu B., Bell R.E., Gerhardt S.P., Manickam J. and Podestà M. 2010 *Phys. Plasmas* **17** 082504

[5] La Haye R.J. 2006 *Phys. Plasmas* **13** 055501

[6] Sabbagh S.A. *et al* 2006 *Nucl. Fusion* **46** 635

[7] Berkery J.W., Sabbagh S.A., Betti R., Bell R.E., Gerhardt S.P., LeBlanc B.P. and Yuh H. 2011 *Phys. Rev. Lett.* **106** 075004

[8] Berkery J.W., Sabbagh S.A., Betti R., Hu B., Bell R.E., Gerhardt S.P., Manickam J. and Tritz K. 2010 *Phys. Rev. Lett.* **104** 035003

[9] Hu B., Betti R. and Manickam J. 2005 *Phys. Plasmas* **12** 057301

[10] Liu Y., Chu M.S., Chapman I.T. and Hender T.C. 2008 *Phys. Plasmas* **15** 112503

[11] Berkery J.W., Liu Y.Q., Wang Z.R., Sabbagh S.A., Logan N.C., Park J.-K., Manickam J. and Betti R. 2014 *Phys. Plasmas* **21** 052505

[12] Berkery J.W., Wang Z.R., Sabbagh S.A., Liu Y.Q., Betti R. and Guazzotto L. 2017 *Phys. Plasmas* **24** 112511

[13] Berkery J.W., Sabbagh S.A., Bell R.E., Gerhardt S.P. and LeBlanc B.P. 2017 *Phys. Plasmas* **24** 056103

[14] Sabbagh S.A. *et al* 2018 Disruption event characterization and forecasting in tokamaks *2018 IAEA Fusion Energy Conf. [EX/P6-26]* (Gandhinagar) (https://conferences.iaea.org/event/151/contributions/5924/attachments/7284/8850/Sabbagh.S-EX-P6-26-Paper-rev2.pdf)

[15] Kaye S.M. *et al* 2019 *Nucl. Fusion* **59** 112007

[16] Strait E.J. *et al* 2019 *Nucl. Fusion* **59** 112012

[17] Moreno R., Vega J., Dormido-Canto S., Pereira A. and Murari A. (JET Contributors) 2016 *Fusion Sci. Technol.* **69** 485

[18] Yokoyama T., Miyoshi Y., Hiwatari R., Isayama A., Matsunaga G., Oyama N., Igarashi Y., Okada M. and Ogawa Y. 2019 *Fusion Eng. Des.* **140** 67

[19] Kates-Harbeck J., Svyatkovskiy A. and Tang W. 2019 *Nature* **568** 526

[20] Rea C., Montes K.J., Pau A., Granetz R.S. and Sauter O. 2020 *Fusion Sci. Technol.* **76** 912–24

[21] Murari A., Rossi R., Peluso E., Lungaroni M., Gaudio P., Gelfusa M., Ratta G. and Vega J. 2020 *Nucl. Fusion* **60** 056003

[22] Pau A., Fanni A., Carcangiu S., Cannas B., Sias G., Murari A. and Rimini F. (the JET Contributors) 2019 *Nucl. Fusion* **59** 106017

[23] Rea C., Montes K.J., Erickson K.G., Granetz R.S. and Tinguely R.A. 2019 *Nucl. Fusion* **59** 096016

[24] Karpatne A., Atluri G., Faghmous J. H., Steinbach M., Banerjee A., Ganguly A., Shekhar S., Samatova N. and Kumar V. 2017 *IEEE Trans. Knowl. Data Eng.* **29** 2318

[25] Raissi M., Perdikaris P. and Karniadakis G.E. 2019 *J. Comput. Phys.* **378** 686

[26] Liu Y., Lao L., Li L. and Turnbull A.D. 2020 *Plasma Phys. Control. Fusion* **62** 045001

[27] Piccione A., Berkery J.W., Sabbagh S.A. and Andreopoulos Y. 2020 *Nucl. Fusion* **60** 046033

[28] Montes K.J., Rea C., Tinguely R.A., Sweeney R., Zhu J. and Granetz R.S. 2021 *Nucl. Fusion* **61** 026022

[29] Mothilal R K, Sharma A and Tan C 2020 (Barcelona, Spain 27–30 January 2020) *Proc. Conf. on F. A. T.* vol 11

[30] Strait E.J., Taylor T.S., Turnbull A.D., Ferron J.R., Lao L.L., Rice B., Sauter O., Thompson S.J. and Wróblewski D. 1995 *Phys. Rev. Lett.* **74** 2483

[31] Sabbagh S.A. *et al* 2006 *Phys. Rev. Lett.* **97** 045004

[32] Berkery J.W., Sabbagh S.A., Bell R.E., Gerhardt S.P., LeBlanc B.P. and Menard J.E. 2015 *Nucl. Fusion* **55** 123007

[33] Bernstein I.B., Frieman E.A., Kruskal M.D. and Kulsrud R.M. 1958 An Energy Principle for Hydromagnetic Stability Problems **244** 17–40

[34] Glasser A.H. 2016 *Phys. Plasmas* **23** 072505

[35] Berkery J.W., Betti R., Sabbagh S.A., Guazzotto L. and Manickam J. 2014 *Phys. Plasmas* **21** 112505

[36] Berkery J.W. *et al* 2014 *Phys. Plasmas* **21** 056112

[37] Boyer M.D. and Chadwick J. 2021 *Nucl. Fusion* **61** 046024

[38] Churchill R.M., Tobias B. and Zhu Y. (the DIII-D team) 2020 *Phys. Plasmas* **27** 062510

[39] Zhu J.X., Rea C., Montes K., Granetz R.S., Sweeney R. and Tinguely R.A. 2021 *Nucl. Fusion* **61** 026007

[40] Zhu M., Xu C. and Wu Y.-F.B. 2013 *Proc. of 13th JCDL* vol 107 (Indianapolis, Indiana, US, 22-26 July 2013)

[41] Hu W.H. *et al* 2021 *Nucl. Fusion* **61** 066034

[42] Lemaître G., Nogueira F. and Aridas C.K. 2017 *J. Mach. Learn. Res.* **18** 1–5

[43] Chen C., Liaw A. and Breiman L. 2004 Using random forest to learn imbalanced data (https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf)

[44] Chawla N.V., Bowyer K.W., Hall L.O. and Kegelmeyer W.P. 2002 *J. Artif. Intell. Res.* **16** 321–57

[45] He H., Yang B., Garcia E.A. and Li S. 2008 1322–8

[46] Bergstra J, Bardenet R, Bengio Y and Kégl B 2011 Advances in Neural Information Processing Systems vol 24 (Granada, Spain, 12–17 December 2011) (https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf)

[47] Montes K.J. *et al* 2019 *Nucl. Fusion* **59** 096015

[48] Zhu W. *et al* 2006 *Phys. Rev. Lett.* **96** 225002

[49] Sabbagh S.A. *et al* 2010 Resistive wall mode stabilization and plasma rotation damping considerations for maintaining high beta plasma discharges in NSTX *Fusion Energy Conf. 2010 (Proc. 23rd Int. Conf.) CD-ROM file [EXS/5–5]* (Daejeon, Vienna, IAEA) (https://pub.iaea.org/MTCD/Meetings/PDFplus/2010/cn180/cn180_papers/exs_5-5.pdf)

[50] Pankin A., McCune D., Andre R., Bateman G. and Kritz A. 2004 *Computer Phys. Comm.* **159** 3

[51] Boyer M.D., Kaye S. and Erickson K. 2019 *Nucl. Fusion* **59** 056008

[52] Gerhardt S.P. *et al* 2012 *Fusion Sci. Technol.* **61** 1

[53] Boyer M.D., Andre R., Gates D.A., Gerhardt S., Goumiri I.R. and Menard J. 2015 *Nucl. Fusion* **55** 053033

[54] Seo J., Na Y.-S., Kim B., Lee C.Y., Park M.S., Park S.J. and Lee Y.H. 2021 *Nucl. Fusion* **61** 106010

[55] Mothilal R K, Mahajan D, Tan C and Sharma A 2021 arXiv:2011.04917

[56] Berkery J.W., Xia G., Sabbagh S.A., Bialek J.M., Wang Z.R., Ham C.J., Thornton A. and Liu Y.Q. 2020 *Plasma Phys. Control. Fusion* **62** 085007

[57] Gerhardt S.P., Andre R. and Menard J.E. 2012 *Nucl. Fusion* **52** 083020