IAEA nuclear fusion

PAPER

Physics-guided machine learning approaches to predict the ideal stability properties of fusion plasmas

To cite this article: A. Piccione *et al* 2020 *Nucl. Fusion* **60** 046033

Nucl. Fusion **60** (2020) 046033 (14pp) https://doi.org/10.1088/1741-4326/ab7597

Physics-guided machine learning approaches to predict the ideal stability properties of fusion plasmas

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 of Great Britain and Northern Ireland
[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 18 December 2019, revised 31 January 2020
Accepted for publication 12 February 2020
Published 18 March 2020

Abstract

One of the biggest challenges to achieve the goal of producing fusion energy in tokamak devices is the necessity of avoiding disruptions of the plasma current due to instabilities. The disruption event characterization and forecasting (DECAF) framework has been developed in this purpose, integrating physics models of many causal events that can lead to a disruption. Two different machine learning approaches are proposed to improve the ideal magnetohydrodynamic (MHD) no-wall limit component of the kinetic stability model included in DECAF. First, a random forest regressor (RFR), was adopted to reproduce the DCON computed change in plasma potential energy without wall effects, $\delta W^{n=1}_{\text{no-wall}}$, for a large database of equilibria from the national spherical torus experiment (NSTX). This tree-based method provides an analysis of the importance of each input feature, giving an insight into the underlying physics phenomena. Secondly, a fully-connected neural network has been trained on sets of calculations with the DCON code, to get an improved closed form equation of the no-wall β limit as a function of the relevant plasma parameters indicated by the RFR. The neural network has been guided by physics theory of ideal MHD in its extension outside the domain of the NSTX experimental data. The estimated value of $\beta^{n=1}_{N,\text{no-wall}}$ has been incorporated into the DECAF kinetic stability model and tested against a set of experimentally stable and unstable discharges. Moreover, the neural network results were used to simulate a real-time stability assessment using only quantities available in real-time. Finally, the portability of the model was investigated, showing encouraging results by testing the NSTX-trained algorithm on the mega ampere spherical tokamak (MAST).

Keywords: disruption prediction, random forests, NSTX, ideal stability, no-wall limit, machine learning, resistive wall mode

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

1. Introduction

High temperature plasmas being studied in tokamak magnetic confinement devices for the purposes of fusion energy must be maintained at high pressures in order to achieve good performance. Unfortunately, at high ratios of plasma pressure to magnetic confinement field pressure, known as β, these plasmas are subject to magnetohydrodynamic (MHD) instabilities that can lead to disruption of the plasma current and impact of the plasma energy on the walls of the device. In order to study and prevent these disruptions, the disruption event characterization and forecasting (DECAF) code [1–3] has been constructed. This code incorporates many different physics models for various causes of disruptions. One of these is the kinetic stability module [1], which directly addresses the issue of high plasma β leading to modes of instability called resistive wall modes (RWMs) and accounts for the stabilizing effects of plasma particle motions [4–6]. Included in this kinetic module is the ideal stability model, which does not account for drift kinetic model effects, but provides the basis for understanding when the plasma has reached a dangerous state and computes quantities that are directly input into the larger kinetic model [7]. The ultimate goal of the DECAF approach is to provide real-time forecasting of plasma stability. Since the computations of both ideal and kinetic stability by traditional means, such as by the DCON [8] and MISK [9] codes, respectively, are too computationally time consuming to be approaches are necessary. For example, efforts have been made to speed up the DCON ideal stability calculation by orders of magnitude [10]. This approach has advantages, as a full MHD calculation can give the stability of any equilibrium solution to the Grad–Shafranov equation, uniquely specified by the plasma boundary shape, as well as pressure and safety factor profiles. In fact, a neural network could also be trained on a large set of stability calculations of model equilibria with these inputs in order to emulate real-time calculations as well. A different approach has been used previously with success: to formulate analytic reduced models for both the ideal [7] and kinetic [1] stability, which maintain the physics basis of the more complex code calculations. In this paper, an extension of that idea is explored, with the help of machine learning (ML).

An important new direction in ML applications is to employ physics-guided, or hybrid, techniques. That is to say, rather than simply feeding raw experimental data into a ML algorithm, physics knowledge of the problem should be utilized to pre-process input into the ML tools, as well as to interpret the output. On the other hand, the patterns discovered through the computations can also be used to improve the physics knowledge, by suggesting previously unappreciated dependencies, for example. In this paper we make the first steps in the direction of this hybrid approach for the specific problem at hand—ideal stability analysis of tokamak plasmas.

ML approaches continue to show outstanding performance in classification tasks. Early-stage neural networks have been proposed as a method for disruption prediction many years ago [11–17] and interest has recently grown rapidly [18–20] with the advent of modern neural network designs that use appropriate non-linearities (i.e. parametric ReLU), customized cost functions for classification problems, and training methods that converge to optima that are shown to be sufficiently stable and surpass conventional optimization methods, like those based on convex approximations. Additionally, neural networks are now being used to model the outputs of plasma physics codes, for example in the area of particle transport [21, 22] or neutral beam injection [23]. Here we use two different algorithms to classify the output of DCON into stable and unstable regions, for a set of calculations for the NSTX tokamak. A large database of equilibria from NSTX has been analyzed with DCON, providing the ideal change in plasma potential energy due to a perturbation of the confining magnetic field without the presence of a conducting wall, $-\delta W^{n=1}_{\text{no-wall}}$. $\delta W^{n=1}_{\text{no-wall}}$ changes from positive to negative when the plasma changes from ideal stable to unstable. We use the negative of the change in potential energy throughout the present paper so that, more intuitively, negative values are *below* the limit (stable) and positive quantities are *above* (unstable). For determining the ideal, and ultimately kinetic mode growth rates, the value of $\beta^{n=1}_{N,\text{no-wall}}$ and both the value and the sign of $-\delta W^{n=1}_{\text{no-wall}}$ must be determined. To that end, we have initially developed a ML based approach to determine $-\delta W^{n=1}_{\text{no-wall}}$ as a function of several plasma parameters. We have tested with both multilayer perceptron (MLP) artificial neural networks and random forest regression (RFR) [24] and we found that the RFR outperformed MLPs in terms of R-squared and flexibility. This model provides an emulation of DCON that could actually run in real-time. The random forest technique has also been recently employed in plasma physics research, specifically in disruption warning, prediction and survival analysis, where a real-time random forest based predictor was employed on the DIII-D, Alcator C-Mod and EAST tokamaks [25–29]. However, no such approach has been attempted for the determination of the ideal stability limit, while also linking the features of this model with respect to their importance. We are proposing such an approach, showing that the original normalized beta, β_N, aspect ratio, A, internal inductance, l_i, and pressure peaking, $p_0/\langle p\rangle$, are among the parameters that most affect the estimated values of $-\delta W^{n=1}_{\text{no-wall}}$, as expected by the underlying physics. Therefore, strengthened by this analysis, we have subsequently used a neural network to obtain an improved equation for the no-wall β limit. Stability regions can be evaluated by examining plots of $-\delta W^{n=1}_{\text{no-wall}}$ in the 2D parameter spaces of β_N versus l_i, β_N versus $p_0/\langle p\rangle$ and β_N versus A. Then neural network defined decision boundaries determine the marginal stability boundaries, increasing the accuracy compared to the previously defined no-wall limit modeling [7]. The overall no-wall beta limit input into the kinetic model can be determined by combining the defined boundaries with the dependencies of $\beta^{n=1}_{N,\text{no-wall}}$ on the parameters above used in the neural network based model.

The remainder of the paper is organized as follows. In section 2 the DECAF code is described, in section 3 we review the physics of a specific module in it which models the ideal global plasma stability. In section 4 the first ML technique, the RFR, is utilized to get estimated values of $-\delta W^{n=1}_{\text{no-wall}}$. Section 5 describes the results of the neural network approach to defining decision boundaries between the stable and unstable regions in plasma parameter space. These are consequently integrated into the DECAF code in section 6, and then applied for cross-machine testing to another spherical tokamak, MAST, in section 7. Finally, section 8 concludes the paper.

2. The DECAF code

Tokamak plasma confinement devices utilize magnetic fields to contain high pressure plasmas for fusion energy. The plasma creates a component of its own confining magnetic field by carrying a large toroidal current. If the current is

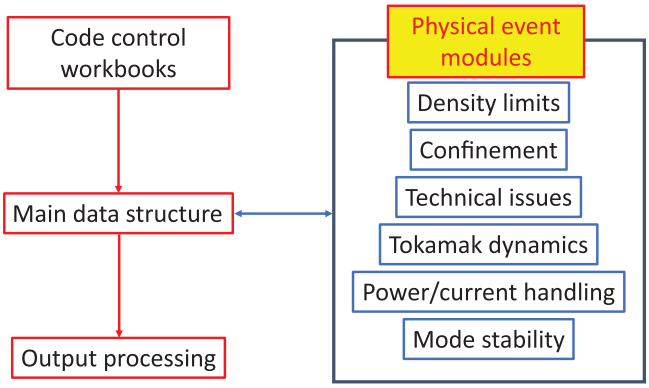


Figure 1. The DECAF framework illustrating physical event modules.

disrupted, a loss of plasma confinement results, which can lead to large heat deposition and electromagnetic forces on the surrounding structures. These so-called *disruptions* [30–32] have varying causes and must be avoided for the safe operation of future fusion devices including ITER. Some of the major causes of disruptions in tokamaks are: when the plasma density, plasma current, or ratio of plasma pressure to magnetic pressure exceed empirical or theoretically defined limits, or when the plasma loses vertical position stability and drifts into contact with surrounding surfaces [33]. A framework has emerged which provides a comprehensive approach to disruption prevention through forecasting and avoidance, or prediction and mitigation of the detrimental consequences. First, it is important to identify disruption event chains and the specific physics elements that comprise those chains. Second, if the events in the disruption chains can be forecast, cues can be provided to an avoidance system to attempt to break the chain. Multiple avoidance actions can be taken through available actuators, giving priority to events that tend to happen earlier. Finally, if avoidance is deemed untenable, a prediction of the impending disruption can be provided to a mitigation system to significantly reduce disruption ramifications. Designs for disruption mitigation systems for ITER are already under way [34–36]. Alternative disruption avoidance approaches also exist, including for example real-time plasma state estimation [37, 38] or by compiling large databases of previous disruptive discharge data [39, 40] and training various ML techniques [41–45] on them, including the previously mentioned neural networks or random forests.

In the present work, we use the DECAF code, which utilizes the comprehensive structure described above. The ultimate goal of such an approach is to provide forecasts which integrate with a disruption avoidance system and are used in real-time during a device's operation. As shown in figure 1, this approach provides a flexible framework to evaluate the proximity of plasma states to detected disruption events by coupled physics analyses, model criteria, and ML techniques.

Although ML approaches are often referred to as *black box* techniques, this definition is actually quite misleading in general. Neural networks have substantial deficiencies when they treat *the system to be modeled* as a black box, rather than making any assumptions about its behaviour, especially when attempting to use resulting neural network matrices for forecasting and extrapolation outside the training domain. In other words, the development of a shallow/deep learning system is complex, but does not need to be a black box. ML in DECAF, by contrast, follows a philosophy that is more amenable to produce human understanding of the results and allows greater flexibility for use in control systems. Specifically, we have presently adopted three general approaches:

(i) Reduction of results from certain complex physical models by shallow ML algorithms to allow rapid (including real time) determination of quantities used in DECAF models.
(ii) *Hybrid model approach* that combines the use of a physical model and ML techniques. The benefits are twofold: on the one hand, ML can be applied to produce the part of the problem that does not have a clear physical model (akin to an *observer* calculation in control theory); on the other, physics insight can be used to regularize a ML algorithm.
(iii) Analysis of DECAF event chain linkages pertaining to disruption prediction and avoidance understanding, utilizing graph theory [46]. This approach can give a mathematical representation of the event chains and can be used to process and interpret data structures in large datasets of DECAF events, in order to gain knowledge on how they group and what are their key relationships.

The analysis presented later in this paper falls mainly into the first category above, while also touching upon the second, which is an important new frontier in ML application to physics problems.

3. Ideal MHD stability in the NSTX spherical tokamak

Global MHD stability of high performance tokamak plasmas has long been recognized as a requirement for fusion-grade applications. Tokamak fusion plasmas are theoretically stable up to a value of the ratio of plasma stored energy to magnetic confining field energy of $\beta_{N,\text{no-wall}}^{n=1}$. With ideal theory, the plasma is unstable above this 'no-wall' limit when no electrically conducting wall is present to external kink-ballooning modes driven by the free energy of current of pressure gradients. Successful wall stabilization of kink/ballooning instabilities uncovered the reduced growth rate, yet still disruptive, resistive wall mode (RWM) [47–50]. The RWM grows on a much slower time scale, the wall-time τ_w, but it is still fast compared to the duration of the plasma shot, and is still theoretically unstable above the no-wall limit with ideal theory. Because the fusion power and the self-generated current in a tokamak rise with β, it is strongly desirable to operate at high β above the no-wall limit. Therefore it is necessary to stabilize the RWM as well. Tokamak experiments found the fortuitous result that plasmas could be stably operated above $\beta_{N,\text{no-wall}}^{n=1}$ [51, 52] with passive stability, not active control. Understanding the physics of this stability is key to relying on it and projecting it to the operation of future devices.

One of the functions of the DECAF code is to monitor fusion plasma's stability and predict when those plasmas might be approaching instability so that something can be done to avoid a catastrophic loss of plasma confinement. In the present paper we will examine data from the national spherical torus experiment (NSTX), which was a spherical tokamak that operated at low aspect ratio and high beta. Disruption of the plasma current leading to thermal and current quenches, halo currents [53, 54], and heat and electromagnetic forces on the device structure were tolerated in NSTX experiments as the energy in the disruptions was not high enough to cause damage to the device. A reduced kinetic model for resistive wall mode stability based upon theoretical understanding has already been included in DECAF [2]. Here we will briefly describe the kinetic model and its underlying ideal stability components.

Calculation of the complex global mode frequency in a plasma, ω, can be performed through the energy principle approach of calculating changes in potential energy (δW) due to various effects and inputting these into a dispersion relation for ω (or stability criteria for the growth rate γ). Hu and Betti derived a modified energy principle for RWMs that includes the kinetic contributions of particle motions [4, 55]. The resulting dispersion relation is:

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

where ω_r is the real frequency of the mode, $\delta W^{n=1}_{\text{no-wall}}$ and $\delta W^{n=1}_{\text{with-wall}}$ are fluid, or ideal, changes in potential energy terms where the stabilizing device conducting structure is omitted or considered, respectively, and δW_K is the kinetic term. Therefore, inclusion of kinetic effects represents a modification to ideal stability [7]. The fluid terms can be broken out into terms depending on the magnetic field line shape, and current-driven and pressure-driven instability terms [56]. This is why measurable quantities such as A, l_i (which is related to current profile peaking), β_N, and $p_0/\langle p\rangle$ are known to be influential on ideal MHD stability. The DCON stability code was developed to calculate the fluid δW terms and was previously used for thousands of calculations spanning the NSTX operating space [7]. Analysis of these thousands of calculations formed the basis of an analytical model for the fluid δW terms that was incorporated into the DECAF code [2]. Improvements to that model now by means of ML is examined in this paper.

ML techniques, guided by physics, will be used as a tool in the present work to determine the best fit to ideal MHD stability limits, as calculated by DCON, within the range of parameters for which experimental data is available. In other words, the DCON calculations *are* the physics guidance within the range of applicability. However, as we will see, ML techniques can also extrapolate outside of the training region, and one must be cautious about accepting those projections at face value. In that case, plasma physics theory of ideal MHD will be used to guide the predictions. For example, it is well known that high pressure peaking is destabilizing to pressure-driven kink modes [57–59], and that Troyon criterion implies that the stability should decrease with aspect ratio [57, 60, 61].

Table 1. List of signals used from NSTX equilibrium reconstructions and diagnostics. Left column shortly describes each signal, while the right column provides the alias name as it appears in the database.

Signal description	Alias
Normalized beta, β_N	betan
Internal inductance, l_i	li
Pressure peaking factor, $p_0/\langle p\rangle$	ppeakfac
Aspect ratio, $A = R_0/a$	aspectratio
Plasma stored energy, W_{MHD}	wmhd
Plasma elongation, κ	kappa
Safety factor at 95% of the flux, q_{95}	q95
Greenwald density fraction, $\bar{n}_e/n_G$	fgw
Toroidal rotation frequency in the core, $\omega_{\Phi(0)}$	ft_core
Toroidal rotation frequency at mid-radius $\omega_{\Phi(\text{mid})}$	ft_mid
Plasma current, I_p	ip
Toroidal beta, β_t	betat
Poloidal beta, β_p	betap
Cylindrical safety factor, $q*$	qstar
Electron density in the core, n_{e0}	ne0
Electron temperature in the core, T_{e0}	te0
Line-averaged electron temperature, $\bar{T}_e$	tebar
Line-averaged electron density, $\bar{n}_e$	nebar
Electron density peaking factor, $n_{e0}/\langle n_e\rangle$	nepeakfac

4. RFR for the value of $-\delta W^{n=1}_{\text{no-wall}}$

4.1. Data pre-processing

Major efforts have been devoted to modeling $-\delta W^{n=1}_{\text{no-wall}}$ and $\beta^{n=1}_{N,\text{no-wall}}$ by defining analytic relationships between plasma parameters [7]. This approach maintains simplicity and causal inference of the dependencies. However, the prediction of the value of $-\delta W^{n=1}_{\text{no-wall}}$ can be improved further by including more plasma parameters and utilizing different approaches. Listed in table 1, are 19 plasma parameters from the equilibrium reconstruction and various diagnostics that can be used as features to predict aspects of MHD stability.

Recent studies have focused on the possibility of combining ML algorithms with prior physics intuition [62–64]. For example, the process of pre-selecting the input parameters reduces the risk of redundancy while keeping physical significance. In this specific application, one can discern from the physics that the first 10 parameters listed above are the ones that are mostly correlated to $\delta W^{n=1}_{\text{no-wall}}$. However, we must first make two important considerations. First of all, the toroidal rotation terms were missing for around 2000 data points and undermined the model performance, as shown later in table 3. Second, when dealing with multivariate regression problems, it is also essential to check that the absolute value of the pairwise correlation between independent variables (predictors) is below a certain threshold, typically 0.7 [65]. Whenever two predictors are highly correlated with one another, both of them will have unstable partial regression coefficients with relatively large standard errors. This issue is usually referred to as *multicollinearity* and can compromise the statistical

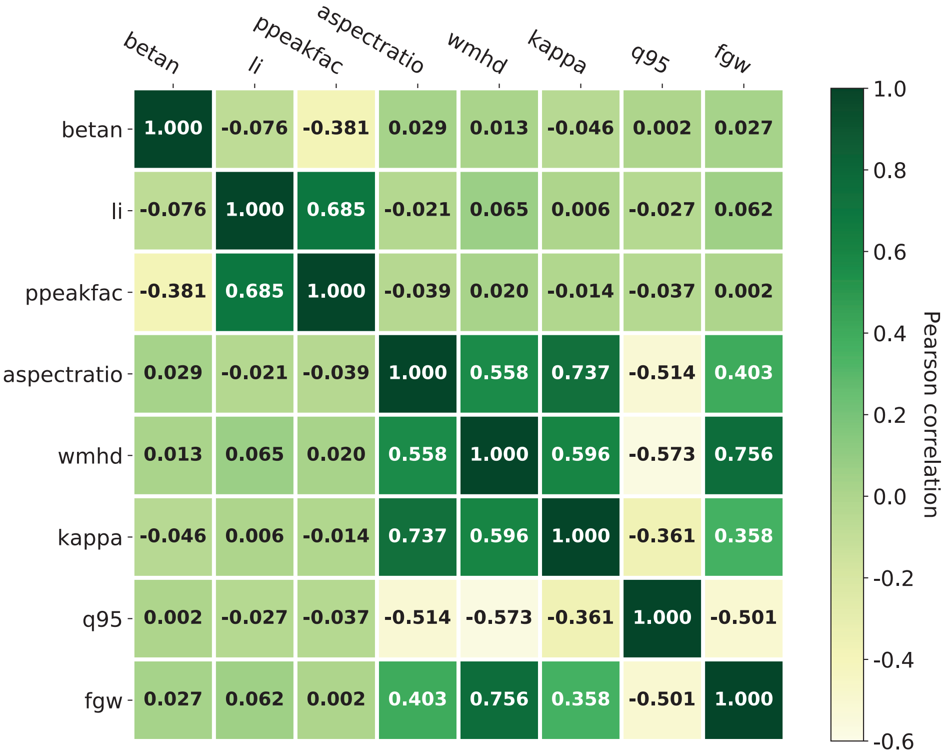


Figure 2. Pearson correlation matrix for the most relevant plasma parameters. A pairwise coefficient above 0.7 generally means that the two predictors are strongly correlated and may affect the model performance.

significance of a predictor, resulting in a higher potential for overfitting.

With regards to this, we have computed the pairwise Pearson correlation between the eight most relevant predictors (excluding rotation terms) as shown in the colormap in figure 2. This analysis shows evidence of multicollinearity between plasma parameters linked to the same underlying phenomena, such as W_{MHD} with $\bar{n}_e/n_G$, and A with κ. Such correlations are expected from constraints related to physical operation of the NSTX device. It is also worth noting that the correlation between pressure peaking factor and internal inductance is quite close to the 0.7 'limit'. This experimental relationship was already known, since pressure broadens with broadened current profiles (see figure 4 in Berkery *et al* [7]). We have found that the two strong correlations introduced a minor overfitting issue, whereas leaving out either $p_0/\langle p\rangle$ or l_i from the feature space led to worse performance, because of the high impact these plasma parameters have on the determination of $-\delta W^{n=1}_{\text{no-wall}}$. The effect of correlated input variables will be analyzed in depth later.

4.2. Hyperparameter tuning

The RFR algorithm belongs to the ensemble methods class, building multiple decision trees and merging them together to get a more accurate and stable prediction. Each tree selects a random number of features in order to better differentiate the influence of each input parameter on the overall predictions. The final output is affected by several *hyperparameters*; among them the most important are the number of trees in the forest, the maximum number of features considered by each tree, the maximum number of levels in each tree, the minimum number of observations to split an internal node and the minimum number of observations to be at a leaf node (deepest level).

The random forest analysis has the advantage of high predictive power for non-linear regression problems and it is regarded as an interpretable technique, since it provides the importance of each input feature and allows to have an insight into decision paths via the `export_graphviz` function in the `scikit-learn` Python library (see figure 4 in Rea *et al* [27]).

We have trained a RFR on a large dataset, comprising of 1385 shots for training and 293 for testing. Each shot contained around five measurements of $\delta W^{n=1}_{\text{no-wall}}$, totalling more than 10 000 data points for the entire set of available equilibria. The RFR has been tuned by running 5-fold cross validation for 600 random combinations of the hyperparameters and for each of the tested models. As widely explained before, the performance of any ML algorithm is strongly affected by the choice of the input variables. The first model utilizing 8 plasma quantities to test the RFR resulted in a minor overfitting problem. Luckily, the RFR algorithm allowed us to analyze the relative importance of each feature (table 2) and to discern which parameters we could sacrifice to reduce the chances of overfitting.

Therefore, based on relative importance, we have decided to drop out κ and $\bar{n}_e/n_G$, solving the overfitting issue without compromising the generalization performance. As an additional test, we have also tried to train the random forest

Table 2. Relative signal importance for the 8-feature RFR. The measure based on which the optimal conditions are chosen is called impurity. For regression trees, this is typically the variance. When training an entire forest, the decrease in impurity due to each feature can be averaged and the features are ranked according to this metric.

Signal	Relative importance
`betan`	0.633 184
`li`	0.200 685
`ppeakfac`	0.087 245
`wmhd`	0.016 737
`fgw`	0.015 968
`aspectratio`	0.015 762
`kappa`	0.015 602
`q95`	0.0148 17

Table 3. Performance of the previous model and the four combinations of plasma parameters input to the RFR, with the best chosen regressor highlighted in bold. The original reduced model was directly validated on the entire dataset, therefore there is no distinction between training and testing set.

	Coeff. of determination (R^2)	
	Training set	Testing set
Original reduced model	—	0.344
RFR—10 features	0.789	0.735
RFR—8 features	0.822	0.776
RFR dropping κ and $\bar{n}_e/n_G$	**0.786**	**0.775**
RFR dropping l_i only	0.621	0.602
RFR dropping $p_0/\langle p\rangle$ only	0.711	0.682

excluding either the pressure peaking factor or the internal inductance. Table 3 shows the results obtained with each of the mentioned combinations in terms of coefficient of determination, R^2, as well as the performance of the original reduced model [7]. Here one can get an idea of the importance that $p_0/\langle p\rangle$ and l_i have in the determination of $\delta W^{n=1}_{\text{no-wall}}$, despite the correlation between the two plasma quantities.

The best selected model had 600 estimators (trees), each one having a maximum of 15 levels. These (and other hyperparameters) were chosen on the basis of the lowest cross-validation mean squared error. The relative importance of the input features is displayed in table 4 and reflects, similarly to the 8-feature analysis, what we expected from the underlying physics.

The `DCON` computed $-\delta W^{n=1}_{\text{no-wall}}$ versus the RFR predicted value as a function of equilibrium quantities is plotted in figure 3. The left hand plot (*a*) displays the results obtained on the training set, whereas the plot on the right (*b*) shows the predictions during the test phase. Each point is colored by the spatial density of nearby points, spanning from dark blue (high density) to yellow (low density). In both cases there is generally a linear correspondence with some spread.

Furthermore, as long as $-\delta W^{n=1}_{\text{no-wall}}$ can take both negative and positive values, we noticed that the stable/unstable classification improves with increasing regression performance. Therefore, we can compute the percentage of misclassified points (e.g. the ones in the upper left and bottom right quadrants of figures 3(*a*) and (*b*)), even though the RFR is performing a regression task. In this case, the RFR classifies points better than the previous model, reaching an overall accuracy of 91.1% in test phase and producing only 2.8% of false negatives and 6.1% of false positives, compared to the original reduced models' 11.2% and 5.6%, respectively.

Table 4. Relative signal importance for the 6-feature regression.

Signal	Relative importance
`betan`	0.688 461
`li`	0.198 174
`ppeakfac`	0.074881
`wmhd`	0.014 078
`aspectratio`	0.013 183
`q95`	0.011 223

Ultimately, we will now proceed to sacrifice performance somewhat in favor of an improved closed form equation for the no-wall limit to be used in the kinetic model.

5. Neural network defined decision boundaries for the no-wall limit

5.1. Base model configuration

The random forest approach can provide the value of $-\delta W^{n=1}_{\text{no-wall}}$ with a relatively high level of accuracy, but complications occurred when trying to use the RFR predictions to get an improved expression of the analytic equation for $\beta^{n=1}_{N,\text{no-wall}}$ already included in `DECAF`. We attempted to train a linear regressor on the input data resulting in RFR predictions of $|\delta W^{n=1}_{\text{no-wall}}| < 0.1$ to find the hyperplane that divides stable and unstable regions, but the two stacked regressors did not provide any improvement when compared to the original reduced model. Another possibility could be to choose a cross section of the 6-dimensional grid and project it down to 2D. However, the decision boundary of a multidimensional model is complex and any 2D projection will not represent the regressor/classifier in its entirety. More accurate Voronoi-based representation methods exist [66], allowing the visualization of multi-dimensional decision regions in 2D. Furthermore, high-dimensional decision boundaries can be characterized by combining adversarial example generation and PCA [67]. However, as a first step we have decided to exploit the knowledge given by the RFR in a different way to re-define an analytic expression for $\beta^{n=1}_{N,\text{no-wall}}$.

A 3-layer fully connected neural network (figure 4) is proposed to classify whether an equilibrium data point from NSTX is below or above the stability limit. The jth layer implements a linear operation that is activated through a logistic function, mapping the input to the subsequent layer (X_{j+1}) into a range between 0 and 1 as follows:

$$X_{j+1} = g(\Theta_j X_j + b_j) \tag{2}$$

where g is the logistic (or sigmoid) activation, Θ_j is the matrix of weights mapping from layer j to $j+1$, b_j is the bias term and X_j is the feature vector input to the jth layer. The dataset has been split in the same way as it was done for the random

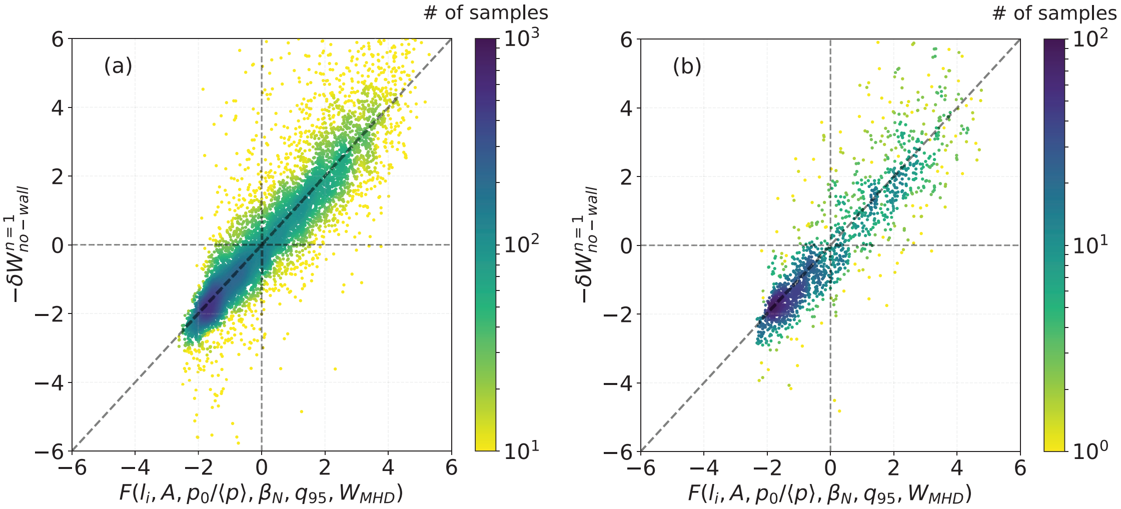
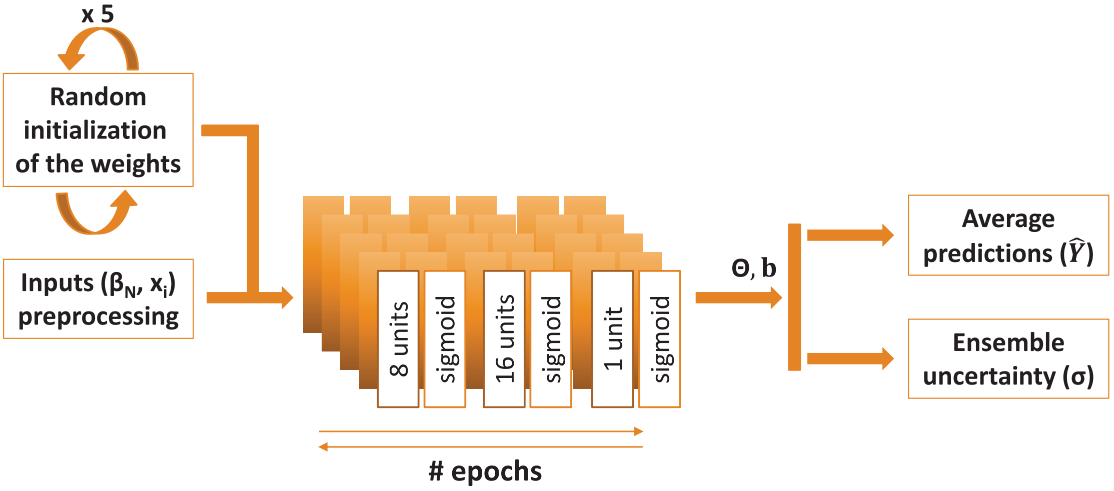


Figure 3. $-\delta W^{n=1}_{\text{no-wall}}$ computed by DCON versus DCON versus $F(l_i, A, p_0/\langle p\rangle, \beta_N, q_{95}, W_{\text{MHD}})$ training and (*b*) testing sets. Each data point is color-coded based on the spatial density of nearby points. Therefore, darker regions indicate higher density, whereas lighter colors indicate higher sparsity of the points.



x 5
Random initialization of the weights
Inputs (β_N, x_i) preprocessing
8 units
sigmoid
16 units
sigmoid
1 unit
sigmoid
# epochs
Θ, b
Average predictions ($\hat{Y}$)
Ensemble uncertainty (σ)

Figure 4. Proposed architecture of the ensemble neural network to classify whether a data point is below or above the no-wall stability limit in the three separate (β_N, x_i) spaces, where x_i is either l_i, $p_0/\langle p\rangle$ or A. The 5 sub-models outputs are amalgamated into an ensemble in order to evaluate the uncertainty on the predictions.

forest for direct comparison. The entire training set is iteratively scanned for 1000 times (epochs) and both the weights matrices and biases are updated at each pass.

We have trained different neural network architectures, varying the number of layers between 2 and 5 and the number of neurons between 8 and 128. The main goal was to find the best compromise between validation accuracy and physical explainability, reducing complexity in order to avoid the risk of overfitting. The learning phase was performed by separately training three identical neural networks with inputs (β_N, l_i), (β_N, $p_0/\langle p\rangle$) and (β_N, A), respectively.

By preselecting these particular plasma quantities we are engaging in the practice of feature engineering, which is to try to feed the predictive model the data most representative of the problem, rather than the most raw data. The features are not only selected to make the most straightforward comparison to previous models, but are also the ones contributing with around 97% of the information according to the random forest. Each feature is min–max normalized since neural networks can better process values between 0 and 1. The network was trained using a stochastic gradient descent optimizer with decaying learning rate in the KERAS library [68] in order to find the optimum combination of parameters that minimizes the binary cross-entropy loss function, $\mathcal{L}$ [69].

Since neural networks can make different decisions depending on how the starting kernels are chosen, we have decided to build an ensemble of five models for each input, each taking a random initialization of the network weights. Similarly to what was done by Boyer *et al* [23], the output is chosen to be the mean of the sub-models predictions and the standard deviation from the mean is utilized to provide a sense of the uncertainty on the predictions. The neural network returns the probability of $-\delta W^{n=1}_{\text{no-wall}} > 0$ (i.e. the probability of belonging to the unstable class). Therefore, points

Table 5. Heaviside step function values used in the penalized objective function. The additional term is not taken into account at low $p_0/\langle p\rangle$, l_i and A.

x_i	$H(x_i)$		
l_i	0	**if**	$0 < l_i < 0.9$
	0.95	**if**	$0.9 \leqslant l_i < 1.2$
	1.2	**if**	$l_i \geqslant 1.2$
$p_0/\langle p\rangle$	0	**if**	$0 < p_0/\langle p\rangle < 3.5$
	0.4	**if**	$3.5 \leqslant p_0/\langle p\rangle < 4.5$
	0.9	**if**	$p_0/\langle p\rangle \geqslant 4.5$
A	0	**if**	$0 < A < 1.5$
	0.2	**if**	$1.5 \leqslant A < 1.6$
	0.4	**if**	$A \geqslant 1.6$

having an output probability <0.5 are considered below the stability limit, while the ones above are classified as likely unstable points. The no-wall limit is defined as the locus of points (here *decision boundary*) where the neural network's outputs are equal to 0.5.

5.2. Physics-guided objective function

We found that if the training phase was performed on a data-driven-only basis, a somewhat surprising result was that the no-wall β limit would monotonically increase with pressure peaking, internal inductance and aspect ratio across the experimental domain. As was noted in section 3, the theoretical expectation is that at higher pressure peaking (and l_i, the two parameters are correlated in NSTX [7]) the no-wall limit should decrease. To a lesser extent, the same is true for the aspect ratio projection, which should also be slightly decreasing rather than the implied slight increase, at higher A. Therefore, the theoretical physics guidance is imposed by modifying the learning objective function [62, 63] for the decision boundary projection at higher $p_0/\langle p\rangle$, l_i and A. Normally, neural networks aim to minimize an empirical loss while keeping a low model complexity for better generalization. However, this approach does not guarantee that the predictions will be in line with the expected underlying physics, especially outside the domain of applicability. Therefore, we introduce an additional term to the loss function which *penalizes* the predictions with increasing $p_0/\langle p\rangle$, l_i and A. Let us denote with $\hat{Y}$ the model predictions and with Y the actual observations, and evaluate the new learning objective as follows:

$$\arg\min_{\Theta,b} \underbrace{\mathcal{L}(\hat{Y},Y)}_{\substack{\text{Empirical}\\ \text{loss}\\ \text{function}}} + \underbrace{\frac{1}{N}\sum_{n=1}^{N}\frac{H(x_i)_n}{\hat{y}_n}}_{\substack{\text{penalization}\\ \text{term}}} \tag{3}$$

where N is the number of training samples and $H(x_i)$ is a two-step Heaviside function that defines the relative importance of the penalization term and it is an additional hyperparameter that needs to be properly chosen for each plasma quantity. In fact, an overly large step function would excessively penalize the predictions, resulting in a too negative steepness of the boundary at high $p_0/\langle p\rangle$, l_i and A, whereas a small H would not provide any change in the predictions. Therefore, we have randomly initialized for 20 times the magnitude of the two steps: the first was chosen in the range [0.05, 2] and the second one was constrained to be larger than the first but less than 4. We have selected the values at which the boundaries smoothly decreased outside of the training region without compromising the validation accuracy within the points. The magnitude of H for each of the plasma parameters is displayed in table 5.

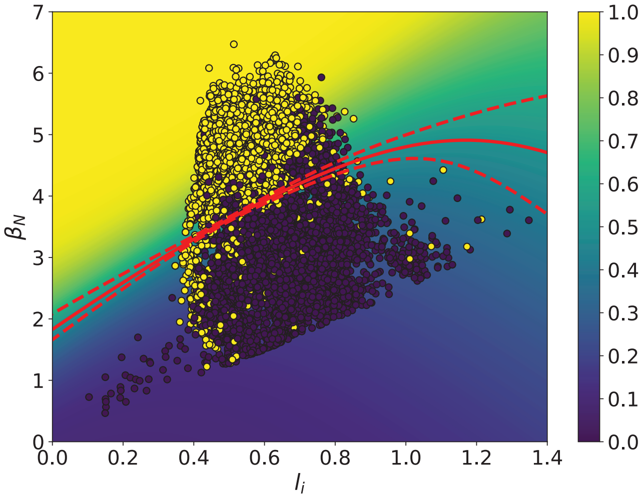


Figure 5. β_N versus l_i space with the red solid line as the mean decision boundary and the dashed lines indicating the standard deviation from the mean. The contour plot shows the probability of being above the no-wall limit.

Stability regions and the uncertainty can be evaluated by plotting β_N versus the other plasma quantities. Figure 5 shows the 2D space defined by β_N versus l_i, with each point representing a DCON calculation for an individual NSTX equilibrium color coded either in purple ($-\delta W^{n=1}_{\text{no-wall}} < 0$, below the 'no-wall' stability limit) or in yellow ($-\delta W^{n=1}_{\text{no-wall}} > 0$, above). The contour plot is what the neural network predicts in each point of the grid in terms of probability. The plots include the physics-guided projections outside of the range where data was available; this point will be discussed further in section 7.

Highlighted as a red solid line is the decision boundary, which effectively defines the location of the no-wall beta limit. The red dashed lines indicate the standard deviation from the mean. As expected, the five sub-models make similar decisions where the density of the points is high, whereas they diverge outside the domain. An analytic expression approximating the no-wall limit can be obtained by fitting the best curve to the boundary points and in this case is given by:

$$\beta_{N,bnd}(l_i) = 4.91\ \mathrm{e}^{-\left(\frac{l_i-1.17}{1.14}\right)^2} + 0.21\ \mathrm{e}^{-\left(\frac{l_i-0.27}{0.39}\right)^2}. \tag{4}$$

Similarly, figures 6 and 7 show the decision boundaries for aspect ratio and pressure peaking, and the resulting equations are:

$$\beta_{N,bnd}(A) = -4.14\,A^3 + 13.47\,A^2 - 14.95\,A + 10 \tag{5}$$

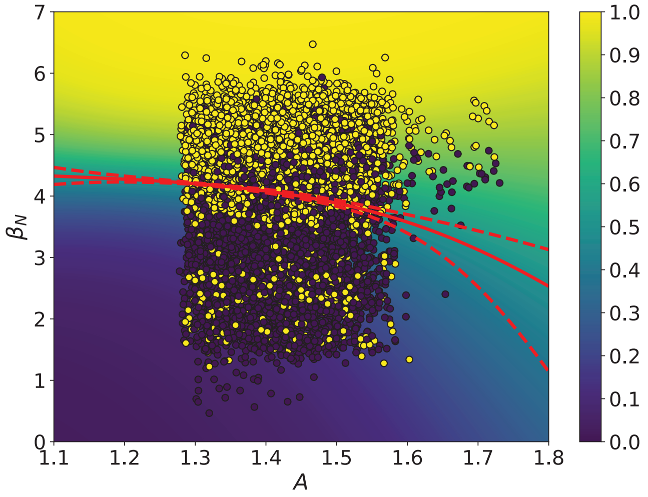


Figure 6. β_N versus A space for the NSTX database. Data points are color coded either in purple ($-\delta W^{n=1}_{\text{no-wall}} < 0$) or yellow ($-\delta W^{n=1}_{\text{no-wall}} > 0.$).

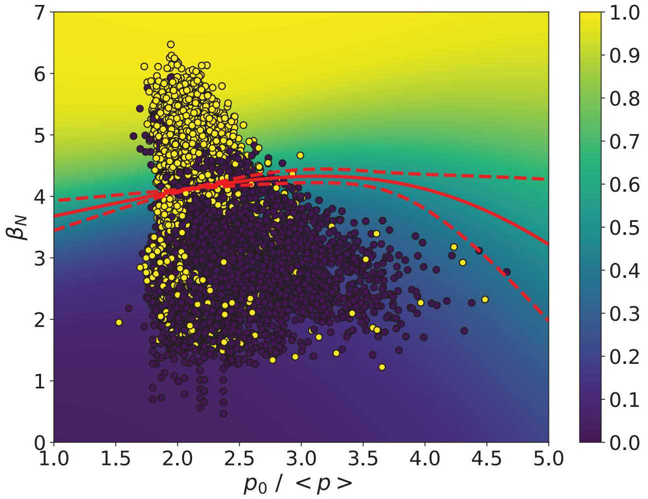


Figure 7. β_N versus $p_0/\langle p\rangle$ space for the NSTX database. Purple is below the no-wall limit (class 0) and yellow is above (class 1).

$$\beta_{N,bnd}\left(\frac{p_0}{\langle p\rangle}\right) = 2.56\,\mathrm{e}^{-\left(\frac{p_0/\langle p\rangle - 4.06}{2.59}\right)^2} + 3.10\,\mathrm{e}^{-\left(\frac{p_0/\langle p\rangle - 0.43}{4.29}\right)^2}. \tag{6}$$

Table 6. Weighting factors and coefficients for the $\delta W^{n=1}_{\text{no-wall}}$ fit.

	Weights	Coefficients
Intercept (a_0)	—	−0.12
β_{N,l_i}	0.3764	3
$\beta_{N,p_0/\langle p\rangle}$	0.3246	1.2
$\beta_{N,A}$	0.2990	1

Table 7. Accuracy for the original splitting equations and for the neural network defined boundaries.

	Accuracy	
	Original (%)	Neural net boundaries (%)
Low l_i and low $p_0/\langle p\rangle$	86.34	87.18
High l_i and high $p_0/\langle p\rangle$	89.47	90.56
Low l_i and high $p_0/\langle p\rangle$	62.58	84.06
High l_i and low $p_0/\langle p\rangle$	71.64	68.63
Overall	**83.16**	**85.92**

5.3. Model performance and uncertainty quantification

The overall no-wall beta limit can be determined by combining these defined boundaries in a way that we will briefly explain. First, by plotting $-\delta W^{n=1}_{\text{no-wall}}$ versus the ratio between β_N and any of the other plasma parameters, it can be easily seen that the best fit is roughly given by a cubic expression (see figure 2(*b*) in Berkery *et al* [7]). The same happens if we plot $-\delta W^{n=1}_{\text{no-wall}}$ versus $\beta_N/\beta_{N,bnd}(x_i)$. Secondly, as long as $\delta W^{n=1}_{\text{no-wall}}$ must be equal to zero at the boundary, the best fit for each of the boundaries should be given by $-\delta W = a_{\beta_N,x_i}$ $(\beta_N/\beta_{N,bnd}(x_i)^3 - 1)$; where the a_{β_N,x_i}'s are coefficients that need to be optimized. Then, each $-\delta W$ term should be appropriately weighted via weighting factors (i.e. w_{β_N,x_i}) provided by the random forest. Therefore, the following fit for $F = -\delta W^{n=1}_{\text{no-wall}}$ will be used:

$$F = a_0 + \beta_N^3 \sum_{x_i} \frac{a_{\beta_N,x_i}\, w_{\beta_N,x_i}}{\beta_{N,bnd}(x_i)^3} - \sum_{x_i} a_{\beta_N,x_i}\, w_{\beta_N,x_i}. \tag{7}$$

The chosen a coefficients are the ones providing the highest R^2 and accuracy. With regards to the weighting factors, one can see that in the above fit the importance of β_N is implicit because it is included in each term. Moreover, since the w_{β_N,x_i} should add up to 1, they need to be rescaled based on the relative importance of l_i, $p_0/\langle p\rangle$ and A from table 4. Both the weights and the coefficients are listed in table 6.

Setting $F = 0$ at the boundary, this results, finally, for an expression for the no-wall beta limit:

$$\beta^{n=1}_{N,\text{no-wall}} = \sqrt[3]{\left[\left(\sum_{x_i} a_{\beta_N,x_i}\, w_{\beta_N,x_i}\right) - a_0\right]\left[\sum_{x_i} \frac{a_{\beta_N,x_i}\, w_{\beta_N,x_i}}{\beta_{N,bnd}(x_i)^3}\right]^{-1}}. \tag{8}$$

We can then propagate the error on the no-wall limit estimation by combining the uncertainty on the decision boundaries with the spread of F around the zero. When the DCON calculated $\delta W^{n=1}_{\text{no-wall}}$ is in the range $[-0.1, 0.1]$, the neural network assisted predictions result in an average of $F_0 = -0.06$ and a standard deviation $\sigma_F = 1.05$. The resulting estimated uncertainty on $\beta^{n=1}_{N,\text{no-wall}}$, $\sigma_{\beta_{N,\text{no-wall}}}$, is obtained using the variance formula for the error propagation:

$$\sigma_{\beta_{N,\text{no-wall}}} = \sqrt{\left(\frac{\partial\beta_N}{\partial F}\,\sigma_F\right)^2 + \sum_{x_i}\left(\frac{\partial\beta_N}{\partial\beta_{N,bnd}(x_i)}\,\sigma_{\beta_{N,bnd}}(x_i)\right)^2} \tag{9}$$

which translates into an error on the estimated no-wall limit of around ±18%.

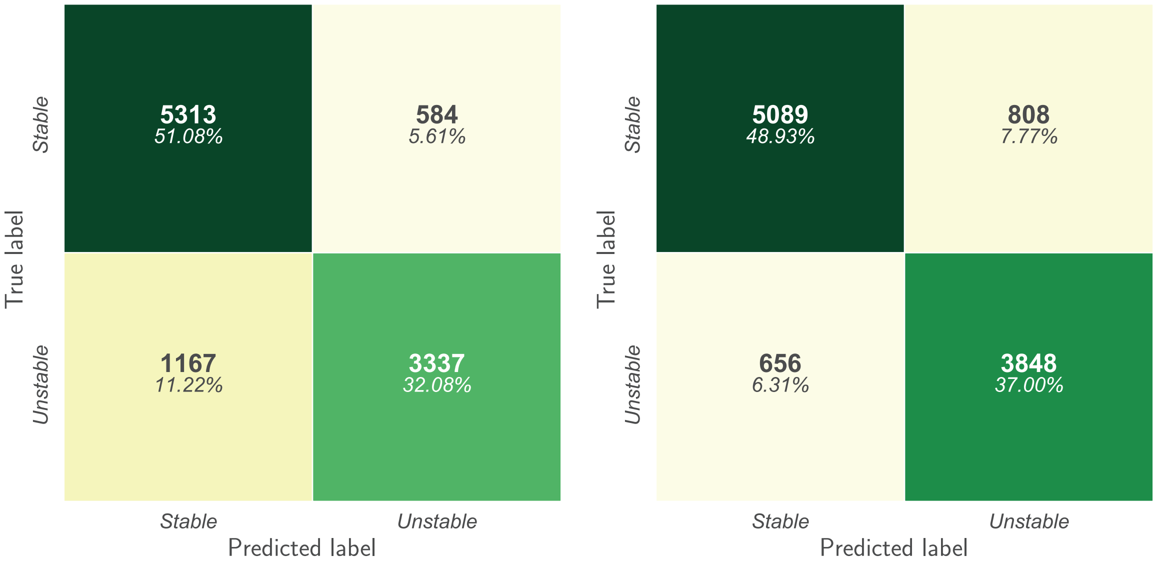


Figure 8. Confusion matrices for the (left) original and (right) neural network decision boundaries no-wall limit. The predictions are split into true negatives (upper-left quadrant), false positives (upper-right), true positives (bottom-right) and false negatives (bottom-left).

The analytic expression obtained has been tested at low $l_i < 0.64$ and high $l_i > 0.64$, as well as at low $p_0/\langle p\rangle < 2.25$ and high $p_0/\langle p\rangle > 2.25$, for best comparison with the previously defined limit. The results are displayed in table 7, where accuracy indicates the percentage of the DCON calculations which fall on the correct side of the stable/unstable boundaries as defined by the original or neural network assisted boundaries. The results show that the neural network outperforms the previously defined equation in Berkery *et al* [7], especially at low l_i, leading to an overall improvement in accuracy of around 2.8%. The overall accuracy can also be displayed in a so-called 'confusion matrix' where the classification is broken into quadrants based on the 'actual' stable/unstable calculation from DCON, and the 'predicted' stable/unstable model based classification.

These are shown in figure 8 for the original and neural network analyses, respectively, where the overall accuracy numbers from the above table are the summation of the diagonal quadrants: calculated stable and predicted stable, and calculated unstable and predicted unstable. Here one can see that there is a slightly larger number of false positives, but with the benefit of an almost halved amount of missed instabilities (i.e. bottom-left quadrant in the confusion matrix). False positives, within certain limits, are acceptable, whilst an abundance of missed instabilities represents a risk for the safety of the reactor.

Overall, the neural network analysis is capable of finding decision boundaries that are still understandable in terms of a few expected plasma parameters and that perform better than the previous analytic technique in terms of classification accuracy. On the other hand, the 2D analysis intrinsically excludes possible correlations between features. The decisions the neural network makes could be stronger and more accurate by feeding it with all the parameters available, rather than just two at a time. However, this approach does not provide a closed form equation for the decision boundary to be included in the larger kinetic model. One possibility could be to use the 19-parameter space shown in table 1 and apply a dimensionality reduction technique (i.e. Principal Component Analysis, t-SNE, etc) to map the original features into a low dimensional space, find the decision boundary in this 'high-level' space and then map it back to real plasma parameters.

6. Integration of ML techniques into the DECAF global stability module

The ultimate goal of the global stability monitoring algorithm is to include not only ideal, but also kinetic effects—which has been shown to be able to explain experimental stability [7]. A reduced model including kinetic effects is already implemented in DECAF, but there are a few possible ways to improve this approach. One would be to utilize ML techniques for the kinetic δW_K terms in the way that we have done here for the ideal no-wall terms, by constructing a neural network or a random forest that can accurately approximate the results of a stability code calculation in a small fraction of time. Unfortunately, this technique works best when many thousands of code calculations are available, spanning the operational space of a device. This is not the case for kinetic calculations. For example, the MISK code has been used to determine the kinetic stability of NSTX for many equilibria, but due to the more demanding computational compared to DCON for example, the number of available MISK runs is certainly not thousands, let alone hundreds. This is why an analytic reduced kinetic model was originally developed [1].

Second, it is possible to run the present analytic reduced model on a very large set of discharges to obtain a database of calculated kinetic growth rates to be compared to experimental stability. Then, a ML technique could replicate that analysis producing stability maps and defining paths for improvement. One key question to be demonstrated by such an approach would be how smoothly the gradients and the partial derivatives flow in the stability map with various parameters. This

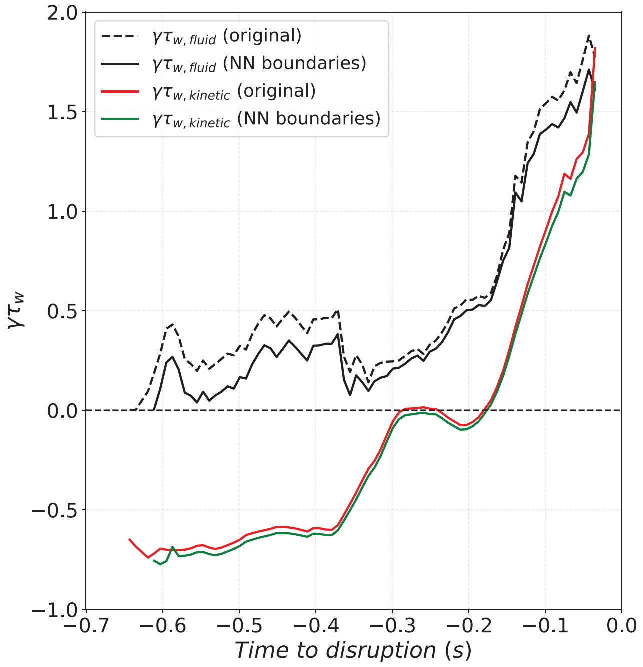


Figure 9. Normalized ideal fluid growth rate versus time to disruption in black dashed line for the original DECAF model, and black solid line using the NN boundaries, for NSTX 139514. The colored lines indicate the original and modified kinetic growth rates in red and green, respectively.

is a necessary line of research for a ML technique that will ultimately be interfaced into a control system. Control algorithms will fail if ML techniques can not provide smooth and reliable stability maps, upon which to act. The research line just described will be the subject of future work.

6.1. Application to post-discharge NSTX data

For the present purposes, a less ambitious approach is to utilize the reduced kinetic model for $\gamma\tau_w$ (as in (1)) currently included in DECAF, but with the neural network improved calculation of the $\beta^{n=1}_{N,\text{no-wall}}$ term (i.e. using (8)). In this case we anticipate small changes, but adding slightly more accuracy, to the computed growth rate.

Figure 9 shows a comparison of the two models for the normalized ideal fluid growth rate in black, and the full kinetic models as colored solid lines versus time leading to disruption, for NSTX shot 139514.

One can see that the solid black line, indicating the neural network modeling of $\gamma\tau_{w,fluid}$, gives smaller values than the current DECAF model (dashed). This is mainly due to the reduced amount of missed instabilities (false negatives) in the present work, which consequently gives slightly higher $\beta^{n=1}_{N,\text{no-wall}}$ values. In fact, since the ideal growth rate is computed as a function of the familiar parameter $C_\beta = (\beta_N - \beta^{n=1}_{N,\text{no-wall}})/(\beta^{n=1}_{N,\text{with-wall}} - \beta^{n=1}_{N,\text{no-wall}})$, the previous observation reflects in lower C_β values, with a smaller $\gamma\tau_{w,\text{fluid}}$ for the entire shot as a direct consequence.

This decrease in $\gamma\tau_{w,fluid}$ propagates in the green line as well, where the kinetic growth rate is shown as a moving average for illustrative purposes. It is worth emphasizing that all the moving averages in DECAF use only previous time points, in order to be applicable in future real-time disruption warning systems. Again, no big changes have been obtained in the final DECAF output, although in this particular case, the improved $\gamma\tau_{w,\text{kinetic}}$ has avoided a possibly erroneous early zero-crossing (i.e. when the model estimates the RWM is going unstable).

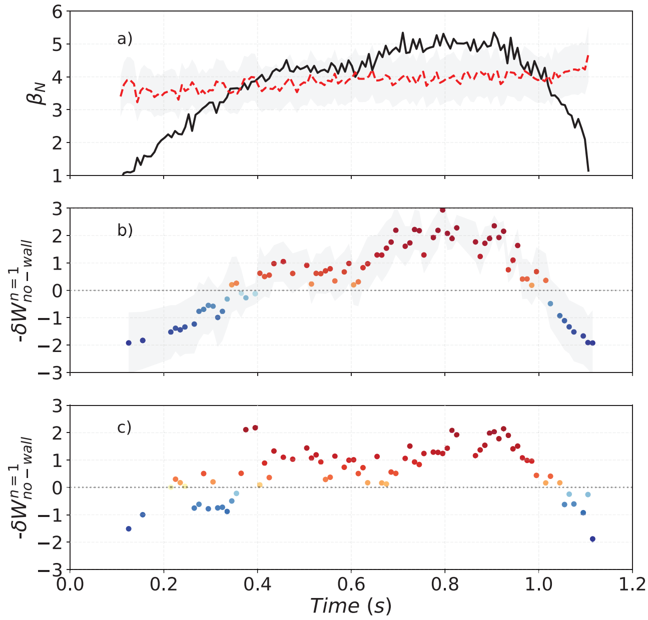


Figure 10. Simulated real-time calculations for NSTX discharge 138556. (*a*) β_N versus time with the estimated no-wall limit and the uncertainty area shown in red and grey, respectively. (*b*) Estimated $-\delta W^{n=1}_{\text{no-wall}}$ and the σ_F as a grey shaded area. (*c*) DCON computed $-\delta W^{n=1}_{\text{no-wall}}$ (ground truth).

When applied to a set of NSTX discharges previously analyzed by the DECAF kinetic model [1], the current physics-guided, ML assisted model failed to predict an unstable RWM in three out of twenty experimentally unstable discharges, whereas just one out of nine stable discharges was predicted to be unstable. The fact that improving one component in the dispersion relation (equation (1)) leads to results consistent with current DECAF calculations is quite promising and lays the basis for further ML assisted computations of the δW_K term.

6.2. Simulated real-time stability calculations

It was previously stated that one benefit of training ML algorithms on databases of calculations is their ability to emulate those calculations much more quickly given the inputs. Tokamaks generally have some measurements and analyses, such as equilibrium reconstruction, available in real-time during the discharge. Real-time disruption avoidance algorithms can take advantage of that by using these inputs and trained ML algorithms to provide stability quantities faster than the real-time operation of the plasma device. In this case we can simulate what the real-time $\beta^{n=1}_{N,\text{no-wall}}$ and $-\delta W^{n=1}_{\text{no-wall}}$ from the neural network would have looked like for an NSTX discharge, using only the inputs that were available in real-time for that shot. Then, we can compare these quantities to the no-wall limit from the DECAF algorithm and $-\delta W^{n=1}_{\text{no-wall}}$ from a DCON calculation, which use post-processed equilibrium reconstructions as input.

Figure 10(*a*) shows in black the real-time equilibrium reconstruction of β_N, with the simulated real-time no-wall limit in red and the $\pm 18\%$ error bar as a grey shaded area, for NSTX discharge 138556. Frame (*b*) displays the emulated $-\delta W^{n=1}_{\text{no-wall}}$ using equation (7) along with the error bars in grey. The DCON computed $-\delta W^{n=1}_{\text{no-wall}}$ versus time from post-discharge analysis is plotted in figure 10(*c*).

The neural network simulation of real-time $\beta^{n=1}_{N,\text{no-wall}}$ and $-\delta W^{n=1}_{\text{no-wall}}$ indicates that the plasma is above the no-wall stability limit for 0.4–1.0 s. The value of a real-time calculation of $\delta W^{n=1}_{\text{no-wall}}$ has been recognized [10], and these alarms will be used in a future DECAF real-time disruption monitoring algorithm. In fact, even if the plasma remained experimentally stable in this particular case, other unintended consequences can occur when a plasma crosses the no-wall limit, such as error field amplification and rotation braking [70]. When combined with other DECAF warnings, these signals can contribute to a more robust real-time disruption warning system. Additionally, these warnings can provide input to a control algorithm which, for example, can apply actuators to maintain β_N at a stable value [71].

7. Cross-device application of ML assisted stability calculations

Future high-powered fusion devices, such as ITER, will necessarily be operated in a much more disruption averse manner than present devices. If ML algorithms for disruption avoidance are to help, they must demonstrate that they can be trained on one device and reliably operated on another. Several efforts are underway to determine if this approach is feasible, including for example cross-machine comparisons of disruption forecasting between the DIII-D and JET tokamaks [20]. Here, our goal is to attempt to use the model trained on NSTX data and apply it to the most similar other device, the spherical tokamak MAST, to discover the advantages and limitations of such an approach. Naturally, the same process as was used here to train a neural network for determination of the no-wall beta limit, or a RFR for emulation of the DCON calculation of $\delta W^{n=1}_{\text{no-wall}}$ can be repeated on the database of MAST equilibria, and it is our intention to do so in the future. However, in order to properly repeat this process, first high quality equilibrium reconstructions using kinetic profiles and motional Stark effect constraints on the q profile for the MAST database are required as the basis of the DCON calculations or as input to the ML algorithms. As these reconstructions are now being generated in present research, we can start with the available magnetics-only equilibrium reconstructions for some examples of cross-machine application. Although this is mostly future work, we will here briefly present some initial examples which show the promise of the approach.

First, we will utilize the NSTX-trained β_N formula on two discharges (25109 and 25112) in MAST, for which the no-wall beta limit was experimentally probed by active MHD spectroscopy [72, 73] and calculated with the MISHKA-1 code [74].

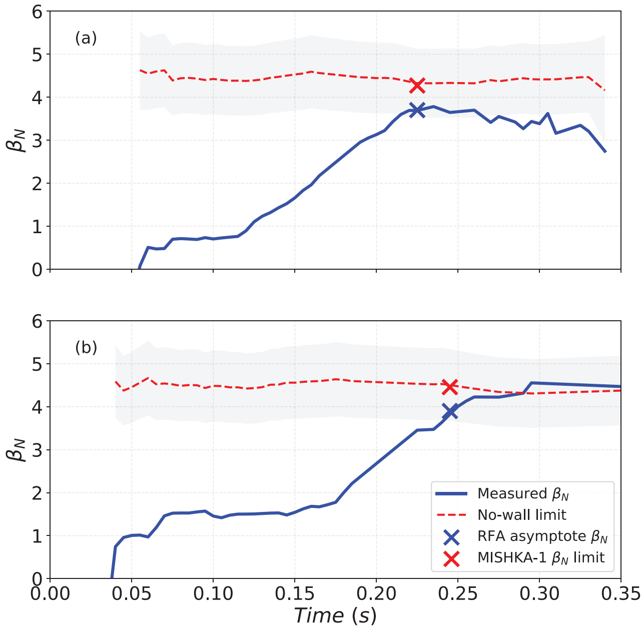


Figure 11. β_N versus time with the neural network no-wall limit trained on NSTX data, for MAST discharges (*a*) 25109 and (*b*) 25112. The grey area indicates the uncertainty on the no-wall limit estimation, as per (9).

Active MHD spectroscopy is an established experimental diagnostic technique used to measure MHD mode stability when the plasma is stable by measuring the resonant field amplification (RFA) of a travelling toroidal mode number $n = 1$ applied tracer field. Experimental evidence to date has shown that increasing amplitude generally indicates decreasing mode stability and a sharply increasing amplitude can indicate the approach to the no-wall limit. By using magnetics-only equilibria and the NSTX derived no-wall limit formula from (8) without any changes for MAST, we can see in figure 11 that the predicted $\beta^{n=1}_{N,\text{no-wall}}$ is somewhat larger than the RFA asymptote, but very close to the MISHKA-1 predictions.

It is worth noting that the operating point of these MAST discharges is just at the edge, or even above, the domain of applicability of the neural network results, with $l_i > 0.8$ and A up to 1.6 at high β_N. Therefore, the NSTX-trained formula applied to MAST has been influenced by the physics guidance outside the training region. This helped to improve the MAST no-wall limit predictions, compared to the previous calculations, as otherwise they would have been even larger. As was stated, this analysis needs to be repeated with accurate kinetic equilibrium reconstructions and DCON calculations of $\beta^{n=1}_{N,\text{no-wall}}$. This example also demonstrates the utility of using multiple machines to validate and refine a ML assisted physics model, as doing so can expand the domain of applicability. This can be tried by training the neural network with just a *glimpse* [20] of MAST data, since this might be the necessary path for future devices such as ITER where a small amount of initial data might be used to update previously trained algorithms, or by combining the full databases of both machines.

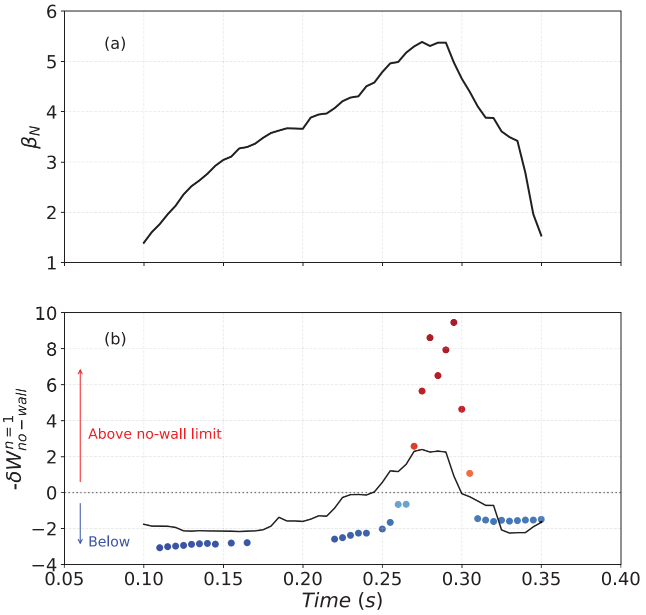


Figure 12. (*a*) β_N versus time and (*b*) $-\delta W^{n=1}_{\text{no-wall}}$ calculations from DCON (colored points), and from the random forest algorithm trained on NSTX data (black line) for MAST discharge 7090.

The second example of cross-machine application from an NSTX-trained ML approach to a MAST discharge is to attempt to emulate DCON calculations of $\delta W^{n=1}_{\text{no-wall}}$. A very high beta MAST discharge, 7090 [75], was selected for this purpose. Figure 12(*a*) shows that β_N peaks in this discharge at ~5.5, and the magnetics-only equilibrium is used as an input to DCON to calculate $-\delta W^{n=1}_{\text{no-wall}}$ in figure 12(*b*). The RFR trained on NSTX data (with kinetic equilibrium reconstructions) was then used to emulate $-\delta W^{n=1}_{\text{no-wall}}$.

The trend is similar, although interestingly the DCON calculations show the no-wall limit (zero crossing) at around 5.2 while β_N is increasing in the shot, and at 4.3 when it is decreasing, whereas the random forest results are more consistent, both times crossing at about $\beta_N \sim 4.6$. The magnitude is also different, although the DCON results of $-\delta W^{n=1}_{\text{no-wall}} > 6$ are unusually large based on the NSTX experience (see figure 3). Overall, $-\delta W^{n=1}_{\text{no-wall}}$ predictions for this MAST discharge, based on a random forest trained on the NSTX device, seem quite reasonable and may be closer in magnitude when compared to potentially new DCON results from kinetic equilibria for MAST. This remains to be seen, but the results are encouraging. Of course, when the random forest predictions are compared to new MAST DCON runs, differences will surely be found as well, which will be informative and in future work may lead to a better predictor.

8. Discussion and conclusions

The reduced kinetic stability model is one of the core modules inside the DECAF framework and has been widely used to reproduce marginal stability points in NSTX. A novel approach combining prior physics knowledge and ML algorithms was explored to improve the underlying ideal stability calculations of the no-wall limit. Interpretable ML approaches, such as the RFR, can give an insight into the physics behind DCON calculations and provide the importance that each plasma quantity has in the determination of $\beta_{N,\text{no-wall}}$ and $\delta W_{\text{no-wall}}$. An improved closed-form equation of the no-wall limit has been derived by combining random forest weights with neural network defined decision boundaries in the β_N versus l_i, A and $p_0/\langle p\rangle$ operating spaces. This new formulation outperforms the previously defined equation, leading to a significant reduction of missed instabilities (false negatives) and consistently replicating DCON calculations in simulated real-time analyses.

There are various elements in this approach that may be improved. First of all, the current DECAF model incorporates modifications to ideal stability by kinetic effects that that will be subject of further ML assisted calculations. Secondly, the usage of neural networks has revealed a caveat of many ML algorithms, which is the extrapolation outside of the training region. Here we have used ideal MHD theory as physics guidance for the neural network. A first cross-device application on a similar tokamak, MAST, is so far quite promising, especially since the neural network has not seen any MAST data. There is still room for improvement as soon as high quality equilibrium reconstructions and DCON runs are available for MAST, though.

In conclusion, a first attempt to include ML tools inside the DECAF framework has proved that physics knowledge and artificial intelligence can cooperate in order to build robust real-time disruption avoidance systems for future relevant fusion devices.

Work in the near future will cover a study of the kinetic effects and the possible procedures to improve the δW_K term, which will provide a ML assisted alternative to the entire kinetic stability model. Moreover, domain adaptation is a necessary line of research in order to refine understanding of how kinetic RWM stability scales across different machines.

Acknowledgments

The authors acknowledge the help of M.D. Boyer, S.P. Gerhardt and R.J. Akers, and are also grateful to A. Abbas for the helpful discussions about customized objective functions.

This research was supported by EPSRC Grants *EP/R025290/1* and *EP/P02243X/1* (University College London), and by U.S. Department of Energy under contract *DE-SC0018623* (Columbia University).

ORCID iDs

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

References

[1] Berkery J.W., Sabbagh S.A., Bell R.E., Gerhardt S.P. and LeBlanc B.P. 2017 *Phys. Plasmas* **24** 056103
[2] Kaye S. M. *et al* 2019 *Nucl. Fusion* **59** 112007
[3] Strait E. J. *et al* 2019 *Nucl. Fusion* **59** 112012
[4] Hu B. and Betti R. 2004 *Phys. Rev. Lett.* **93** 105002

[5] 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

[6] Sabbagh S.A. *et al* 2010 *Nucl. Fusion* **50** 025020

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

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

[9] 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

[10] Glasser A.S., Kolemen E. and Glasser A.H. 2018 *Phys. Plasmas* **25** 032507

[11] Hernandez J.V., Vannucci A., Tajima T., Lin Z., Horton W. and McCool S.C. 1996 *Nucl. Fusion* **36** 1009

[12] Wroblewski D., Jahns G.L. and Leuer J.A. 1997 *Nucl. Fusion* **37** 725

[13] Vannucci A., Oliveira K.A. and Tajima T. 1999 *Nucl. Fusion* **39** 255

[14] Pautasso G. *et al* 2002 *Nucl. Fusion* **42** 100

[15] Cannas B., Fanni A., Marongiu E. and Sonato P. 2004 *Nucl. Fusion* **44** 68

[16] Windsor C.G., Pautasso G., Tichmann C., Buttery R.J., Hender T.C., JET EFDA Contributors and The ASDEX Upgrade Team 2005 *Nucl. Fusion* **45** 337

[17] Yoshino R. 2005 *Nucl. Fusion* **45** 1232

[18] Ferreira D.R., Carvalho P.J., Fernades H. and JET Contributors 2018 *Fusion Sci. Technol.* **74** 47

[19] Zheng W. *et al* 2018 *Nucl. Fusion* **58** 056016

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

[21] Citrin J., Breton S., Felici F., Imbeaux F., Aniel T., Artaud J.F., Baiocchi B., Bourdelle C., Camenen Y. and Garcia J. 2015 *Nucl. Fusion* **55** 092001

[22] Meneghini O. *et al* 2017 *Nucl. Fusion* **57** 086034

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

[24] Breiman L. 2001 *Mach. Learn.* **45** 5

[25] Rea C., Granetz R.S., Montes K.J., Tinguely R.A., Eidietis N., Hanson J.M. and Sammuli B. 2018 *Plasma Phys. Control. Fusion* **60** 084004

[26] Rea C. and Granetz R.S. 2018 *Fusion Sci. Technol.* **74** 89

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

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

[29] Tinguely R.A., Montes K.J., Rea C., Sweeney R. and Granetz R.S. 2019 *Plasma Phys. Control. Fusion* **61** 095009

[30] Hender T.C. *et al* 2007 *Nucl. Fusion* **47** S128

[31] Boozer A.H. 2012 *Phys. Plasmas* **2012** 058101

[32] Eidietis N.W. *et al* 2015 *Nucl. Fusion* **55** 063030

[33] Aydemir A.Y., Lee H.H., Lee S.G., Seol J., Park B.H. and In Y.K. 2016 *Nucl. Fusion* **56** 054001

[34] Hollmann E.M. *et al* 2015 *Phys. Plasmas* **22** 021802

[35] Baylor L.R. *et al* 2015 *Fusion Sci. Technol.* **68** 211

[36] de Vries P.C., Pautasso G., Humphreys D., Lehnen M., Maruyama S., Snipes J.A., Vergara A. and Zabeo L. 2016 *Fusion Sci. Technol.* **69** 471

[37] Geelen P., Felici F., Merle A. and Sauter O. 2015 *Plasma Phys. Control. Fusion* **57** 125008

[38] Blanken T.C., Felici F., Galberti C., Vu N.M.T., Kong M., Sauter O., de Baar M.R., The EUROfusion MST1 Team and The TCV Team 2019 *Nucl. Fusion* **59** 026017

[39] Pau A. *et al* 2017 *Fusion Eng. Des.* **125** 139

[40] Bo W., Granetz R., Bingjia X., Jiangang L., Fei Y., Junjun L. and Dalong C. 2016 *Plasma Sci. Technol.* **18** 1162

[41] Murari A., Vega J., Rattá G.A., Vagliasindi G., Johnson M.F., Hong S.H. and JET-EFDA Contributors 2009 *Nucl. Fusion* **49** 055028

[42] Vega J., Dormido-Canto S., López J.M., Murari A., Ramírez J.M., Moreno R., Ruiz M., Alves D., Felton R. and JET-EFDA Contributors 2013 *Fusion Eng. Des.* **88** 1228

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

[44] 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

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

[46] Cranmer M. D., Rui X., Battaglia P. and Ho S. 2019 (arXiv:1909.05862v2)

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

[48] Sabbagh S.A. *et al* 2002 *Phys. Plasmas* **9** 2085

[49] Sabbagh S.A., Bell R.E., Menard J.E., Gates D.A., Sontag A.C., Bialek J.M., LeBlanc B.P., Levinton F.M., Tritz K. and Yuh H. 2006 *Phys. Rev. Lett.* **97** 045004

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

[51] 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

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

[53] Gerhardt S.P., Menard J.E., Sabbagh S.A. and Scotti F. 2012 *Nucl. Fusion* **52** 063005

[54] Gerhardt S.P. 2013 *Nucl. Fusion* **53** 023005

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

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

[57] Menard J.E., Jardin S.C., Kaye S.M., Kessel C.E. and Manickam J. 1997 *Nucl. Fusion* **37** 595

[58] Turnbull A.D., Taylor T.S., Chu M.S., Miller R.L. and Lin-Liu Y.R. 1998 *Nucl. Fusion* **38** 1467

[59] Huysmans G.T.A. *et al* 1999 *Nucl. Fusion* **39** 1489

[60] Troyon F., Gruber R., Saurenmann H., Semenzato S. and Succi S. 1984 *Plasma Phys. Control. Fusion* **26** 209

[61] Menard J.E. *et al* 2004 *Phys. Plasmas* **11** 639

[62] 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

[63] Karpatne A., Watkins W., Read J. and Kumar V. 2017 (arXiv:1710.11431v2)

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

[65] Dormann C.F. *et al* 2013 *Ecography* **36** 27

[66] Migut M.A., Worring M. and Veenman C.J. 2015 *Data Mining Knowl. Disc.* **29** 273

[67] Karimi H., Derr T. and Tang J. 2019 (arXiv:1912.11460)

[68] Chollet F. 2018 *Deep Learning with Python* (Shelter Island, NY: Manning Publications)

[69] Goodfellow I., Bengio Y. and Courville A. 2016 *Deep Learning* (Cambridge, MA: MIT Press)

[70] Piovesan P. *et al* 2017 *Plasma Phys. Control. Fusion* **59** 014027

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

[72] Chapman I.T. *et al* 2011 *Nucl. Fusion* **51** 073040

[73] Chapman I.T., Gryaznevich M.P., Howell D.F., Liu Y.Q. and The MAST Team 2011 *Plasma Phys. Control. Fusion* **53** 065022

[74] Mikhailovskii A.B., Huysmans G.T.A., Sharapov S.E. and Kerner W. 1997 *Plasma Phys. Rep.* **23** 844

[75] Hole M.J. *et al* 2005 *Plasma Phys. Control. Fusion* **47** 581