RESEARCH ARTICLE | MARCH 02 2026

Interpreting AI for fusion: An application to plasma profile analysis for tearing mode stability

Special Collection: Papers from the 5th International Conference on Data-Driven Plasma Science

Hiro J. Farre-Kaga ; Andrew Rothstein ; Rohit Sonker; SangKyeun Kim ; Ricardo Shousha ; Minseok Kim ; Keith Erickson; Jeff Schneider; Egemen Kolemen

Phys. Plasmas 33, 032502 (2026)
https://doi.org/10.1063/5.0311201

Interpreting AI for fusion: An application to plasma profile analysis for tearing mode stability

Cite as: Phys. Plasmas 33, 032502 (2026); doi: 10.1063/5.0311201
Submitted: 7 November 2025 · Accepted: 7 February 2026 ·
Published Online: 2 March 2026

View Online Export Citation CrossMark

Hiro J. Farre-Kaga,[1,2,a)] Andrew Rothstein,[1] Rohit Sonker,[3] SangKyeun Kim,[2] Ricardo Shousha,[2] Minseok Kim,[1] Keith Erickson,[2] Jeff Schneider,[3] and Egemen Kolemen[1,2,b)]

AFFILIATIONS

[1]Princeton University, Princeton, New Jersey 08540, USA
[2]Princeton Plasma Physics Laboratory, Princeton, New Jersey 08540, USA
[3]Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA

Note: This paper is part of the Special Topic on Papers from the 5th International Conference on Data-Driven Plasma Science.
[a)]**Author to whom correspondence should be addressed:** hf8585@princeton.edu
[b)]**Electronic mail:** ekolemen@pppl.gov

ABSTRACT

Artificial intelligence models have demonstrated strong predictive capabilities for various instabilities in fusion devices such as Tokamaks, including tearing modes (TM), edge localized modes, and disruptive events, but their opaque nature raises concerns about safety and trustworthiness when applied to fusion power plants. Here, we present a physics-based interpretation framework using a TM prediction model as a demonstration that is validated through a dedicated DIII-D TM avoidance experiment. By applying Shapley analysis, we identify how profiles such as rotation, temperature, and density contribute to the model's prediction of TM stability. Our analysis shows that in our experimental scenario, core electron temperature and rotation peaking play the primary role in TM stability, while density changes have smaller effects on stability. We show that off-axis ion temperature stabilizes TMs, suggesting that off-axis neutral beam heating can further stabilize this scenario. This work presents a generalizable ML-based event prediction methodology, from training to physics-driven interpretation, bridging the gap between physics understanding and opaque ML models.

I. INTRODUCTION

Tokamaks are a promising fusion energy technology, but they face challenges in maintaining plasma stability. Recently, machine learning (ML) and artificial intelligence (AI) have been applied more and more to the field of nuclear fusion in the form of surrogate physics models,[1–4] event prediction models,[5–10] and reinforcement learning controllers.[11,12] However, a key requirement for next-generation fusion power plants is to have interpretable control systems where causes of control actions can be directly tied to observations in the plasma. This is directly at odds with the typical AI approach that utilizes the high prediction accuracy of black-box models that are uninterpretable.

Disruption prediction approaches have utilized "gray-box" models, such as random forest models,[8] that offer interpretable results at the trade-off of simpler ML architectures with lower accuracies. However, simple black-box models such as multilayer perceptrons (MLPs) have the advantage of ease of training, allowing researchers to produce highly accurate machine learning models with fewer resources and expertise. Interpretable neural networks may require restricting the model's architecture and number of parameters, which can lead to lower accuracy than a deep complex network.[13]

Instead of changing the "black-box" model architectures to gain interpretability, we can adjust our analysis framework to gain insights from these "black-box" models using Shapley analysis. Shapley analysis is a method for explaining the output of machine learning models based on a game theoretic approach[14] by fairly distributing the prediction result across model inputs. This analysis framework can be applied to any model, machine learning-based or otherwise.

In this application, we study how the plasma profiles affect tearing mode (TM) stability and explain which specific profile features, such as rotation peaking, can stabilize TMs. Previous TM prediction and stability models used only scalar parameters or magnetic field features,[15–18] including a Shapley analysis in Ref. 19 or full plasma profiles with no stability interpretation.[6] We improve on these with an ML-based Deep Survival Machine model[20] to predict TMs based on real-time plasma profiles. Using Shapley analysis, we can explore how the plasma profiles affect TM stability and explain what specific profile

features, such as higher core T_e and T_i, led to the avoidance of TMs. This model is applied to a dedicated TM avoidance experiment on DIII-D, and its results are used for this analysis. While Shapley analysis has been applied to other fields[21] with some applications to fusion,[22,23] it has not been used for interpreting experiments directly as far as the authors are aware.

TM stability analysis poses an interesting problem for interpretation, as many physics studies have been performed to better understand how these instabilities onset[24] and what scalar parameters are most important,[25] often with conflicting results or limited to scenario-specific operating regimes such as the DIII-D ITER baseline.[26–28] Other approaches to better understand TM stability involve using a physics code, such as STRIDE, to calculate the classical Δ' stability parameter;[29] however, this has not been validated in experiments. The state of the art for neoclassical TM theory is the Modified Rutherford Equation (MRE),[30] which determines the island growth rate from a modified Δ' to include the perturbed neoclassical current drive,[31] radial transport within the island,[32] and polarization current.[33] According to MRE, if a "seed" island occurs that is large enough to overcome the stabilizing terms, a neoclassical TM appears; however, the free parameters in MRE make it difficult to use for experimental prediction and database verification. By studying the TM prediction model with Shapley analysis, which has database-wide predictive ability, this can add additional information to the greater plasma physics discussion of TM stability.

This paper begins with an explanation of the tools and techniques used for training and interpreting the TM predictors in Sec. II. Section III describes the TM predictor model results, followed by an interpretation for the TM preemptive avoidance experiment. Then we use Shapley values to draw conclusions more broadly about profile-based TM stability. Finally, Sec. IV summarizes our findings and describes the future work to improve TM models and their interpretation.

II. METHOD

We begin with a description of the database processing used to train our TM prediction model in Sec. II A, followed by an explanation of the training method in Sec. II B and the Shapley model interpretation in Sec. II C.

A. Database processing and TM labeling

The model was trained on all DIII-D shots identified to have the required data between shots 140 000–195 000, resulting in 6050 shots of which 1476 contained n = 1 TMs, where n is the toroidal mode number, which amounts to 677 494 timesteps each with 42 parameters. The data needed is listed in Table I and includes Thomson Scattering,[34] Charge Exchange Recombination Spectroscopy,[35] Motional Stark Effect,[36] magnetics for EFIT reconstructions,[37] and actuation values such as neutral beam and electron cyclotron heating power. While training scenario-specific models may improve performance in that scenario, the aim of this model is to be flexible so it may be applied to different scenarios for DIII-D experiments.

Importantly, our dataset has no differentiation between classical TMs and neo-classical TMs (NTMs) since they both appear similarly in our automated labeling. However, the planned control actions should be effective for both TM and NTM stabilization by replacing the missing bootstrap current. Consequently, when we refer to TMs, it is assumed to include both classical and neo-classical TMs. We also do not consider locked or quasi-stationary modes and limit our analysis to born-rotating modes, as locked modes are often preceded by born-rotating modes.[38]

TABLE I. TM prediction model input parameters and their corresponding sources. Two models are presented: The large model, which uses all the inputs listed above, and the reduced model containing only the plasma profiles (inputs above the horizontal line). EFITRT2 is a real-time magnetic equilibrium reconstruction using magnetic diagnostics and motional Stark effect.

Input	Source
Electron temperature profile (T_e)	RTCAKENN
Electron density profile (n_e)	RTCAKENN
Ion temperature profile (T_i)	RTCAKENN
Ion rotation profile (Rot)	RTCAKENN
Plasma pressure profile (p)	RTCAKENN
Safety factor profile (q)	RTCAKENN
Current density profile (j)	RTCAKENN
NBI power	Neutral beam injection
NBI torque	Neutral beam injection
ECH power	Electron cyclotron heating
I_P	Plasma current
B_T	Toroidal magnetic field
Normalized pressure (β_n)	EFITRT2
$q_{\min}$	EFITRT2
Internal inductance (l_i)	EFITRT2
Plasma minor radius (a_{minor})	EFITRT2
Plasma major radius (R)	EFITRT2
Bottom triangularity (δ_{bot})	EFITRT2
Top triangularity (δ_{top})	EFITRT2
Elongation (κ)	EFITRT2
Plasma volume (Vol)	EFITRT2

The following are the key database processing steps taken and their rationale, with further details covered in Appendix A:

- Plasma current (I_P) rampup and rampdown are excluded as we are only targeting TMs in I_P flat-top.
- A TM was considered to have occurred if the amplitude of the root mean square (RMS) of the n=1 magnetic fluctuations stayed above 12 G for 50 ms, along with additional constraints on H_{98} and q_{95} to only consider H-mode plasmas. Once a TM is identified, its onset time is the time when the n=1 RMS amplitude first reaches 10% of the peak n=1 RMS amplitude.
- Magnetic fluctuation signals, such as the n = 1, 2, 3 magnetic RMS signals that identify the n = 1, 2, 3 modes, are excluded from model inputs in the training set to avoid the model overly relying on these signals as they are used for labeling.
- The data are taken every 20 ms, as this is enough time for diagnostics and EFITs to yield updated results, faster than τ_R and τ_E ensuring the profiles are equilibrated, but not too fast that the model overfits to noise. Actuation, such as changes to ECH deposition and NBI power adjustment, will also affect the plasma on the order of 100 ms, so this is a good compromise.

B. The survival regression training scheme

The model in this paper uses the Deep Survival Machines (DSM) architecture from the open-source `Auton`-Survival package.[20] The framework allows for easy-to-use event prediction, and has been demonstrated in fusion applications for disruption prediction[5] to achieve longer warning times compared to other models. Like any event prediction model, the two key ingredients are accurate labels and input data that are representative of the underlying physics. Figure 1 depicts the training scheme, where in training, we input the plasma parameters, such as the profiles, the time until the end of the sequence, and whether or not a TM occurs at the end of the sequence. At inference, or when running the model, we only input the plasma parameters since the event or time-to-event is not known. This training setup enables a single-time step event to be included in model training without arbitrarily chosen time-to-event ramps, and allows the inference model to use only present data for real-time applicability.

The DSM architecture uses Survival Regression, a statistical scheme that provides the probability of an event occurring at any time within a user-chosen time horizon, $t_{horizon}$, given a set of input features. The model itself consists of an MLP that produces a series of survival distributions, which are combined to produce a final survival probability distribution.

A common application of this algorithm is in estimating the survival times of patients given certain treatments and symptoms, hence the name survival. By analogy, a TM in a plasma may be considered a "death," and the input features, such as the density and temperature, are the "symptoms." This is therefore applicable to plasma disruptions in general, and to specific MHD events such as the onset of a $m, n = 2/1$ TM for poloidal mode number m and toroidal mode number n.

The uses of survival models and the `Auton` framework for this application are well-motivated for the use of right-censored data and the handling of multiple competing risks. We generate our TM labels as standard right-censored data, where we consider a time of event that is either a TM onset or the end of plasma current flattop with no TM onset. This is ideal because a time slice 100 ms before the end of plasma flattop is inherently different from a time slice that is 100 ms before a TM onset. This also enables the model to learn exceptionally stable plasmas and know that a time slice 1 s before the end of the flattop will be stable with high certainty, while another time slice 100 ms before the end of the flattop is more unknown. Additionally, TM stability is a non-linear process and so will have many competing risks. For example, higher β_N values will make TMs more unstable, but also increase plasma rotation, which should stabilize the plasma. Because these cannot be fully decoupled, it is ideal to utilize the `Auton` framework that generates an intermediate representation to systematically handle these unknown competing risks.

C. Shapley analysis for model interpretation

Machine learning models such as the above survival regression are often considered black boxes, as they consist of matrix multiplication sequences with millions of uninterpretable parameters. While these matrix weights are difficult to justify, we can still understand and explain a black-box model by studying how the input features affect the output. For example, a sudden drop in rotation may lead to a TM, so such an input change should increase the model's TM probability. Similarly, this analysis provides an insight into *why* the model predicts a TM, and what plasma feature was most responsible.

Shapley analysis is a game theoretic approach[14] to analyze the impact of input parameters on the model output. Shapley values represent the contribution of a particular feature value to the overall prediction of the model relative to a background distribution. The Shapley value corresponding to an input represents its contribution to the model output relative to the average effect in the background distribution. Hence, the total Shapley value equals the model's output minus the model's average output in the background distribution. For example, if a certain time slice has a Shapley value of 0.1 on core electron temperature, that core electron temperature contributes to a 10% increase in TM probability. To demonstrate this technique, we have provided an illustrative toy example in Appendix C.

In our Shapley analysis TM study, we use the 11 shots from our dedicated DIII-D experiment as the background distribution. These shots are never-before seen for the TM prediction model, as the model was trained on historical data prior to real-time deployment. Another choice for the background would be the entire DIII-D dataset, but narrowing this down to the experiment shots allows for a more in-depth comparison between the 11 shots. For example, β_N will not be a significant factor in shots from our experimental session that achieved similar $\beta_N \sim 3$, but if we were to compare to standard DIII-D H-mode shots with $\beta_N \sim 2$, the effect of β_N would be overwhelming. Thus, in our evaluation of TM predictions later, it is important to ground these interpretations based on the relevant reference distribution of plasma equilibria.

III. MODEL PERFORMANCE AND PHYSICS INTERPRETATION

This section describes the application of the TM prediction model, from its performance statistics in the DIII-D TM database to the model's application to a dedicated control experiment, and finally, an analysis of the plasma profile features impacting TM stability. We begin listing the input features to the model and presenting its performance metrics in Sec. III A. The DIII-D experiment on preemptive TM avoidance is explained in detail in Sec. III B, showing successful TM avoidance and detailing the model's results shot-by-shot. We look into specific time slice profiles and discuss the interpretation of the

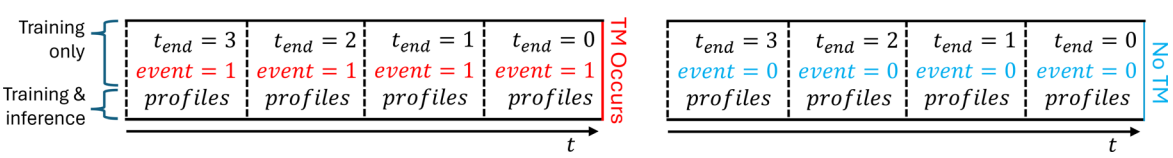

FIG. 1. Depiction of the survival regression training scheme. In training, the model is input t_{end} representing the time until the end of the sequence, *event* representing whether a TM occurs at the end of the sequence (1) or not (0), and *profiles* representing the set of diagnostics and inputs to the model. At inference, only *profiles* are input, since the end of the sequence or event is, of course, not known.

model using Shapley analysis in Sec. III C, where the actuator effects on equilibrium profiles are shown to stabilize TMs. Finally, in Sec. III D, we study the Shapley values across our experiment to draw broader conclusions on the scenario's stability.

A. Tearing mode prediction model

A TM prediction model, referred to as the "Large Model," was trained to predict DIII-D n = 1 TMs using the parameters shown in Table I, including the real-time kinetic profiles RTCAKENN,[39] an ML surrogate model for CAKE,[40] as well as external heating and actuation, and EFITRT2 scalars. The seven RTCAKENN profiles are each reduced to four principal component analysis (PCA) components to reduce the input size with minimal reduction in information, with a reconstruction R^2 value of 0.9991. This led to an input size of 42. The decisions on database selection and processing are listed in A. Most decisions were made to fit the experimental design, which required a real-time capable flattop TM predictor with enough warning time to change the electron cyclotron heating (ECH) deposition and affect the equilibrium (around 200 ms).

The basic performance metrics are shown in Fig. 2, demonstrating high performance metrics with warning times around 1000 ms and an AUROC score of 0.85. This allows for flexibility in actuation to avoid the oncoming TM, such as ECH deposition in the case of our experiment. Throughout the 1000 ms time horizon used in this model, future actuation is assumed constant, as future actuation cannot be used in a real-time application. While this limits the model's predictive ability due to unforeseen TM-triggering actuation, the majority of DIII-D shots in flattop have near-constant actuation, so TMs can be predicted in the future. Further notes on the definition of warning times and classification details are given in Appendix B.

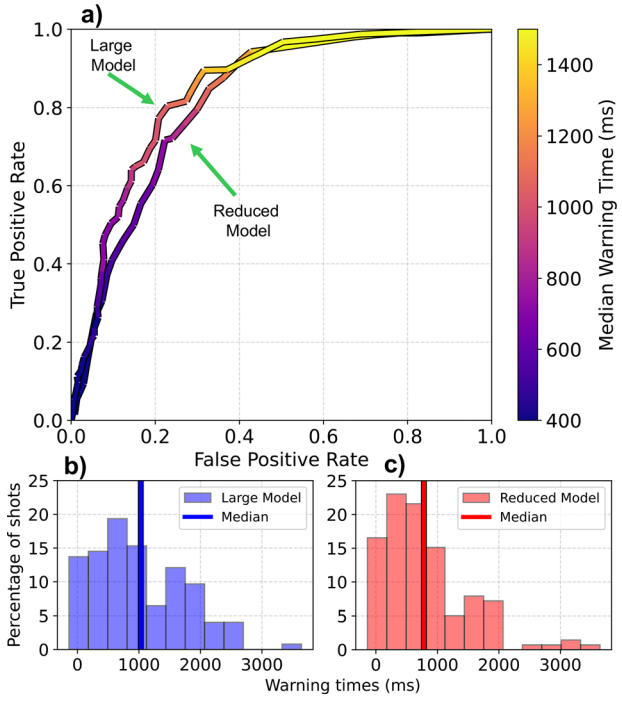

FIG. 2. (a) ROC curve for the RTCAKENN-based TM predictor used in the experiment. The original model had an AUROC score of 0.85, compared to 0.82 for the profiles-only model. (b) The warning time's histogram of true positives for a typical threshold of 0.2 shows that the majority of TMs are predicted over 500 ms in advance, allowing for flexibility in actuation.

The model architecture was a Deep Survival Machine with the following parameters: a 80/10/10 train/validation/test split, an MLP with four hidden layers, each with a width of 128 nodes, a lognormal distribution, a batch size of 256, learning rate of 2×10^{-4}, $k = 2$ survival distributions, dropout of 0.4 and 100 epochs. These parameters were the result of a grid search hyperparameter optimization. The train/validation/test split was done on a shot by shot basis, instead of time slice by time slice, to avoid inflating performance metrics where potentially similar timeslices from a single shot could end up in all three groups.

A second "Reduced Model" was trained using the same database as the "Large Model," but with reduced inputs consisting solely of the seven 33-point RTCAKENN profiles in Table I. This model is used in Sec. III C to study how profile changes affect the TM risk without confounding scalar inputs. For example, β_N, input heating and some shape parameters were found to have consistent, large Shapley values, which made analysis of profile importance more difficult. In addition, PCA components of profiles are not used here to preserve localized features. This model, of course, has worse performance metrics as it used fewer diagnostics and inputs, but the AUROC score of 0.82 is not significantly worse for the purpose of the Shapley analysis. A difference in warning time distribution between the two models can be seen in Figs. 2(b) and 2(c). This model was an MLP with three hidden layers, each with a width of 512 nodes, a log-normal distribution, a batch size of 512, learning rate of 3×10^{-5}, $k = 3$ survival distributions, dropout of 0.7 and 100 epochs.

B. DIII-D experiment: Preemptive tearing mode suppression

The TM predictor was developed for a dedicated DIII-D experiment to achieve active 2/1 TM suppression, consisting of aiming electron cyclotron current drive (ECCD) at the $q = 2$ surface when a TM risk was predicted. Whenever ECH is applied in this experiment, our experimental setup provided ECCD as well. We therefore refer to the combined effect as ECH and differentiate these when referring to just the heating or current drive effect. The experiment was run in the elevated q_{min} scenario,[41] a high-performance advanced non-inductive scenario that is often limited by TMs.

Previous experiments have shown the potential of preemptive suppression of TMs using ECCD[42] by changing the ECCD deposition when TMs are detected. However, 2/1 modes are difficult to suppress once they have appeared and cause significant performance degradation while they are present. We therefore designed a preemptive scheme with the results shown in Fig. 3, where we predict TMs before they appear, and change the ECH deposition as TM probability increases, enabling fully tearing-free operation. As three ECH gyrotrons were available, increasing thresholds of 0.1, 0.2, and 0.3 TM probability were chosen to successively increase current drive at q = 2 with increasing TM risk. The experiment successfully demonstrated the suppression scheme, as can be seen in Fig. 3, resulting in the tearing-free operation of previously unstable conditions.

Importantly for this paper, the TM model successfully predicted the TMs with sufficient warning time to enable actuation and responded correctly to the stabilizing effects of modifying ECH deposition. The experiment started at shot 199 597, which was a reference elevated q_{min} shot, and ended at 199 607. The reference was well predicted, as shown in Fig. 3, followed by another correct unstable prediction in 199 598 and 199 599. Shots 199 605, 199 606, and 199 607 were the same unstable conditions with predicted TMs, but the active adjustments of ECH deposition successfully avoided TMs. Shots 199 600 and 199 601 used additional ECH power, which led to passively stable conditions, correctly predicted again. The only failure of the predictive model was shots 199 602 and 199 603, which were run at lower plasma current, creating n = 3 modes that triggered n = 1 TMs.

The experiment resulted in a 82% success rate, where the two failures were difficult for our model to predict as higher n modes drove n = 1 modes, something our current model is unable to account for, as larger n modes do not have significant profile flattening effects. Future predictor models could incorporate information from magnetic fluctuation signals to have information about higher n modes and their triggering of n = 1 modes.

C. Shapley analysis I: What drove the TM, and how was it suppressed?

Shapley analysis is performed on the two key shots of our experiment shown in Fig. 3: shot 199 597 as the no-control baseline, where a TM occurred and was predicted 500 ms in advance; and shot 199 606, which had the same actuation and conditions as the reference, with the sole difference that ECH deposition is changed from an off-axis current drive position to the q = 2 surface whenever TMs were predicted.

For this analysis, we refer to the calculated Shapley value, which is the percent increase that an input has on the TM probability, as "TM impact" to make the interpretation of the values explicitly clear. In Figs. 4 and 5, a positive TM impact (positive Shapley value) describes TM destabilization and is represented by redder colors, while a negative TM impact (negative Shapley value) causes TM stabilization and is represented by bluer colors. The green vertical bar represents the $q = 2$ surface location, where the tearing mode appears, according to EFIT02ER[43] at $t = 3500$ s. We use the Reduced Model whose inputs were solely the seven profiles defined in Table I, which explains the small differences in TM probability predictions for the same shots between Figs. 3, 4, and 5. These shots are never-before seen by the model, and are at the edge of the training dataset distribution because similar elevated q_{min} shots exist in the database, but without real-time actuation. We aim to understand how the equilibrium profiles determine the stability of the plasma; therefore, the actuation scalars were not included in the model as they indirectly impact the stability by changing the profiles. While Shapley analysis may be performed on the original model, it is important to remove highly correlated values, such as β_N, those that are correlated with the pressure profile, to avoid ambiguity in Shapley values. Using only profiles enables a study of their true impact on TM stability. Error analysis of the TM impacts displayed here is shown in Appendix D.

The four profiles types shown in Fig. 4 reflect the evolution of the plasma and its impact on TM risk. The general increase in T_e leads to more positive TM impact, particularly in the off-axis region ($0.3 < \psi_N < 0.8$). The pedestal growth in the edge region ($0.8 < \psi_N$) has a stabilizing effect seen with the negative TM impact values. This is generally consistent with the understanding that higher temperatures and pressures will lead to more tearing risk, due to a higher β_n and bootstrap current.[44]

Similarly, the evolution of TM impact values for the rotation profile provides insight into the causes of the observed TM. As the core rotation increases, the region becomes more stabilizing as would be expected from an MHD stability perspective. The higher edge rotation has the opposite, slightly destabilizing effect, which suggests that rotation gradient or peaking is an important feature. This is a common experimental observation and is reported in a multi-machine study.[45] The net change in the rotation profile's effect on TM risk (shown in the figure legend) from t_1 to t_3 is a stabilizing 0.1, which is lower than the 0.15 destabilizing impact of the T_e increase, suggesting that this TM is primarily caused by high core T_e.

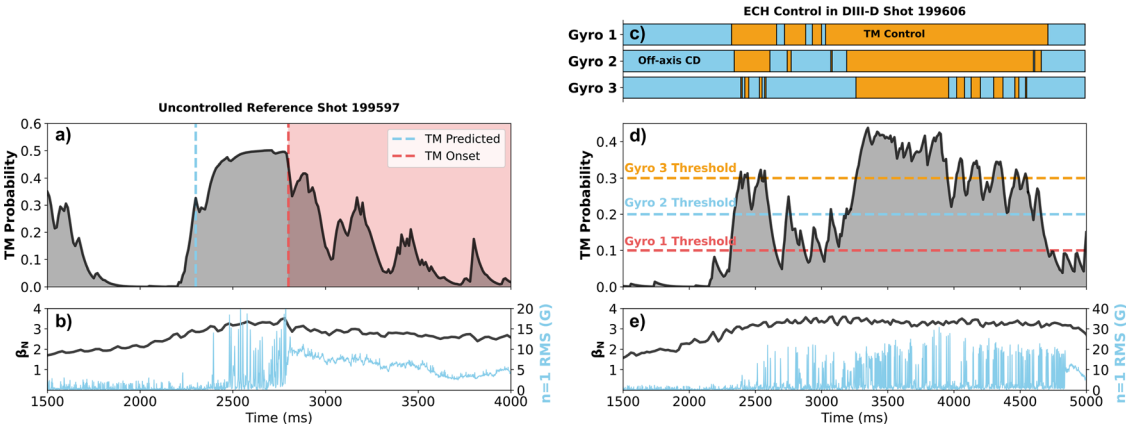

FIG. 3. Demonstration of active preemptive TM suppression via ECH deposition change. (a) and (b) show the uncontrolled reference shot, which resulted in a TM at t = 2700 ms, predicted at t = 2300 ms. (c), (d), and (e) show a shot with TM suppression via ECH deposition change. In (c), blue represents ECH aimed at an off-axis location and orange represents ECH aimed at the q = 2 surface for TM control. As the TM probability increases and the thresholds are crossed, more ECH gyrotrons are steered to the q = 2 surface to stabilize.

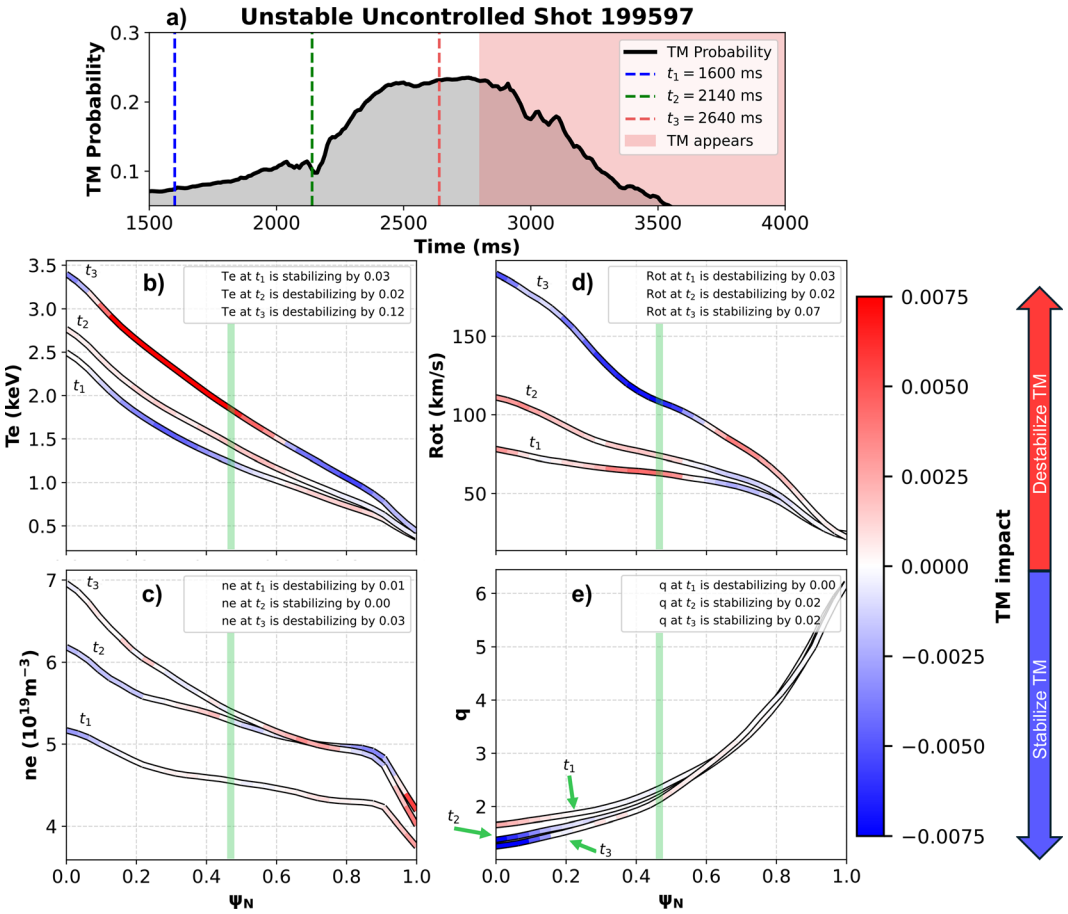

FIG. 4. Shapley analysis for the unstable uncontrolled reference shot 199 597. Three timesteps are chosen, $t_1 = 1600$ ms at the stable start of the shot, $t_2 = 2140$ ms right before the TM is predicted, and $t_3 = 2640$ ms where the tearing risk peaks and a TM occurs at $t = 2700$ ms. The color of the profiles represents the TM impact or Shapley value, where red indicates destabilizing (positive TM impact), blue indicates a stabilizing (negative TM impact), and white indicates no impact. The total impact on stability of the profile, or the sum of Shapley values, is in the legend. Abbreviations are used from Table I.

The density profile evolution has a relatively low impact to TM probability, but has interesting features. The electron density profile is elevated in t_2 and t_3 compared to t_1 even though t_3 is more destabilizing than t_1 or t_2. This suggests the peaked shape in t_3 is responsible for the destabilizing effect rather than the elevated average densities. This is well supported by theoretical studies,[46] which show the density profile only has a slight effect on TM stability and finds density gradients to be destabilizing. Finally, while the q profile evolution had little influence on tearing, the figure is included, as it shows interesting physics information. The drop in q_{min} has a stabilizing effect on the plasma, but the $q = 2$ region notably becomes an unstable red for t_3 relative to t_1, which may be interpreted as the region being at a location with a tearing risk.

A key question we seek to answer with Shapley analysis is the impact of ECH deposition changes at the $q = 2$ surface on TM stability. Despite most actuation being equal to the reference shot in Fig. 4, Fig. 5 shows a control shot where a TM does not appear despite it being predicted by the model, likely due to the real-time modifications to ECH deposition, which increased the current drive at the $q = 2$ surface. In the control shot in Fig. 5, the two timesteps chosen are t_1, where the tearing risk is high, but the ECCD has not yet driven much current in the $q = 2$ surface, and t_2, where the ECCD has provided a more current drive aimed at the $q = 2$ surface and consequently causing the TM probability to drop.

Several profile effects are expected from changing the ECH deposition to target the $q = 2$ surface. Primarily, we expect a changed electron temperature due to heating, as seen in Fig. 5(b). Because our ECH setup also provides current drive, the $q = 2$ region has a bump in current density, j. Finally, ECH has a density pumpout effect, which may explain the lower density at the t_2 off-axis and edge regions, although other factors will impact this too.

The increase in T_e between t_1 and t_2 is subtle, but it causes a small destabilizing effect, as was observed for the reference shot. However, the increase in J_{ECCD} at the $q = 2$ surface, causing the bump on j at t_2 has a small stabilizing effect seen in Fig. 5(d), as is intended from driving ECCD at the $q = 2$ surface where 2/1 TMs appear. Finally, the lower density in Fig. 5(c) has an overall stabilizing effect, particularly in the core and edge regions.

While the two q profiles in Fig. 5(e) are too similar to draw conclusions on their shape, the overall stability of the profile drops by 0.06 after ECCD stabilization of the shot. Since the q profile largely

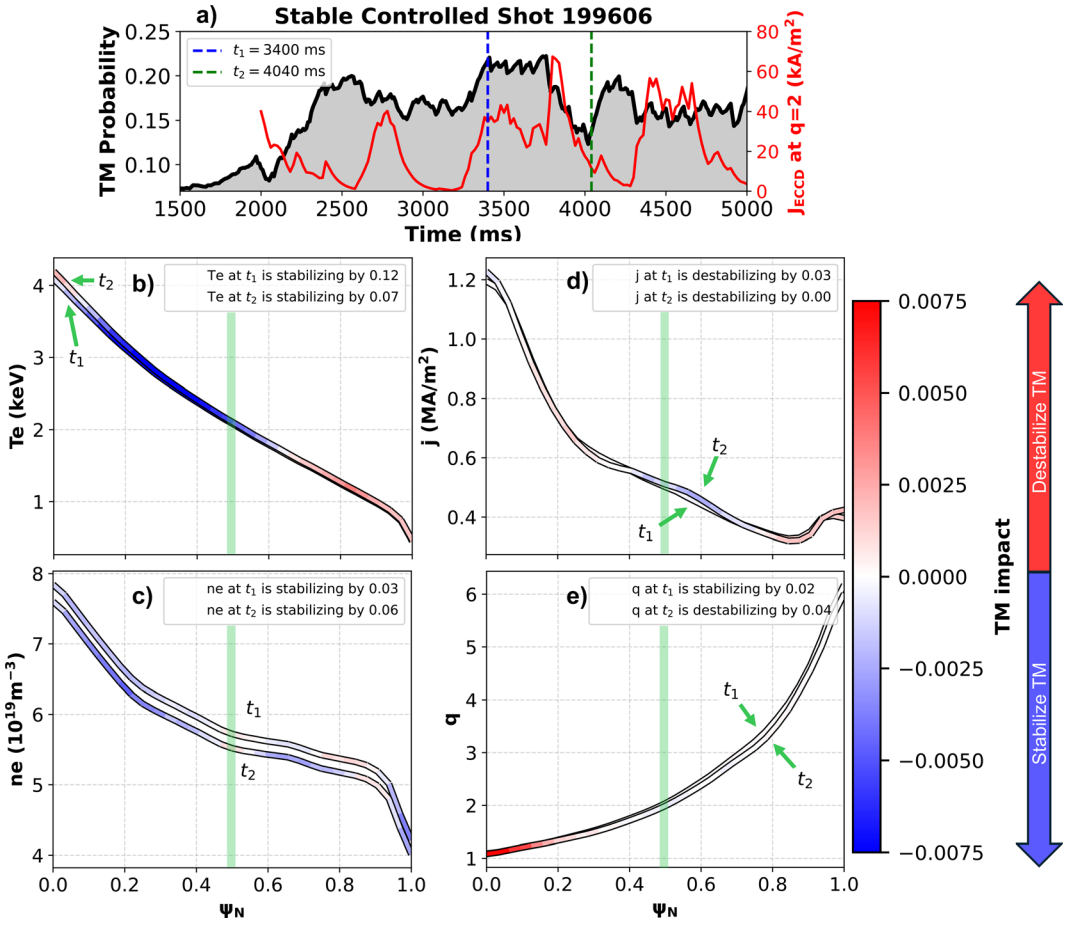

FIG. 5. Shapley analysis for the ECH-controlled stable shot 199 606. Two timesteps are chosen, $t_1 = 3140$ ms when J_{ECCD} at $q = 2$ is low, and the tearing risk is high, and $t_2 = 3920$ ms where ECH deposition has been changed, so J_{ECCD} at $q = 2$ is high, and the tearing risk begins to drop. The rotation profile is not shown, as the small difference between t_1 and t_2 had little impact.

describes the scenario, this could suggest the shot is evolving into a tearing unstable regime, but is kept stable by other profile changes induced by the ECH deposition change.

Overall, we observe that the three key profile changes from ECH deposition modification, namely, localized current drive, electron heating, and density pumpout, have an important effect on TM stability, suggesting that changes in ECH deposition played a role in avoiding the TM in this shot. This analysis shows the insight that Shapley analysis can have on the underlying physics learned by machine learning models for TMs. It also provides information on the triggering mechanisms and causes of a TM, such as a rise in temperature and pressure while rotation remains low.

D. Shapley analysis II: Which profile features affect TM stability?

Shapley analysis can be applied to a wider database analysis, providing insight into the overall impact of a profile feature, to draw more generalized conclusions. In this section, we study which profile features have the largest impact in the scenario of our dedicated experiment.

In Fig. 6, we plot the histograms of TM impact for each profile feature, with the "core" region spanning $\psi_N \in [0, 0.3]$, "off-axis" region being $\psi_N \in [0.3, 0.8]$ and "edge" being $\psi_N \in [0.8, 1]$ to represent the pedestal. The features are ordered by their overall TM impact, specifically by the mean of the absolute value of each TM impact, meaning the feature can be strongly stabilizing or destabilizing to TMs. This is visualized by the width of the histograms in the figure. By this metric, core T_e is the highest and most important for TM stability prediction. It clearly shows that the higher the core T_e value (or the lighter the color), the more destabilizing it will be to TMs. The same pattern is observed for core T_i, off-axis T_e, and core q to a lesser extent. Notably, edge T_e and T_i display the opposite behavior, with higher pedestal temperatures contributing to lower TM risk.

The profiles largely derived from magnetic measurements, j and q, have generally smaller TM impacts. This is likely due to all shots covered in this analysis having similar j and q profiles, as they are in the same scenario. With less variation in j and q profiles, we would expect a smaller TM impact from these profiles. However, their core values show a clear pattern of high j and therefore low q leads to more TM stable shots, which is the expected result as stronger current drive

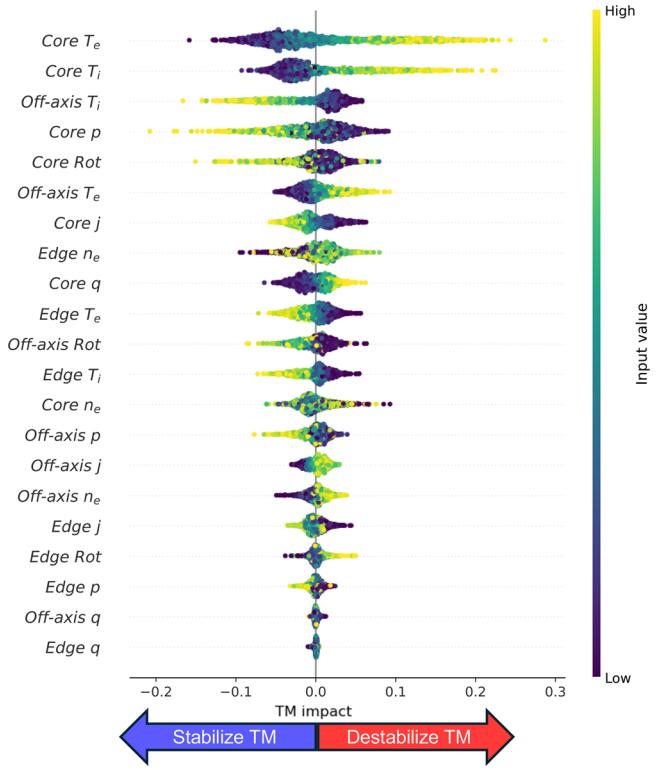

FIG. 6. A general Shapley analysis of all inputs, ordered by influence on TM stability predictions in the dedicated experiment. Each row is a histogram of TM impact for a given value, with its color indicating the magnitude of the input. For example, core T_e has the widest distribution, suggesting high importance for TM prediction. As core T_e increases (color becomes yellower), its TM impact clearly increases, showing that a large core T_e has a strong destabilizing effect.

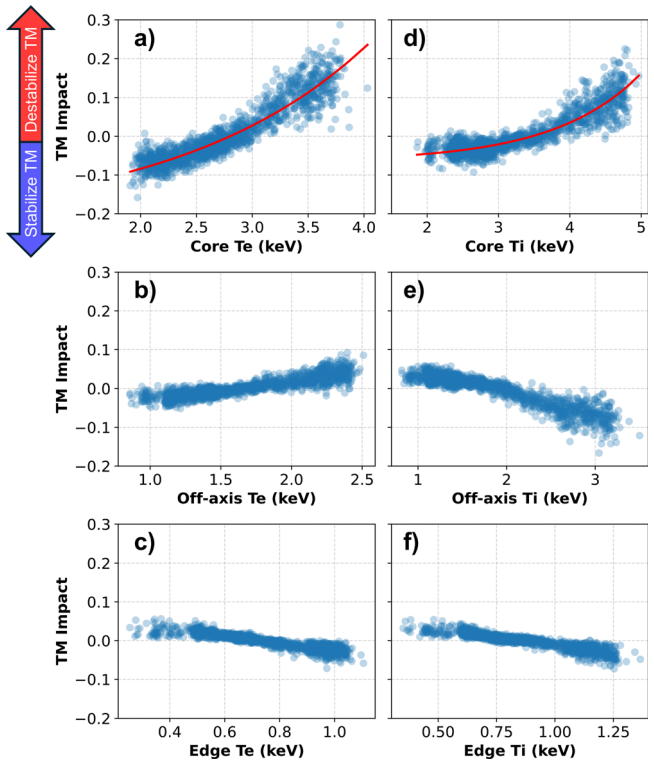

FIG. 7. Scatter plots of TM impact from T_e and T_i across core, off-axis and edge regions, with each point representing one time slice in the experimental data. From these plots, we can draw conclusions on the stability impact of different profile regions. The red curve is an exponential fit to guide the eye.

should reduce TM risk. Real-time magnetic reconstructions used in this analysis also have lower accuracy, which may impact these results.

The core and off-axis rotation are strongly stabilizing, as is expected from large flows suppressing MHD instabilities. However, the core rotation distribution in Fig. 6 has a dark center with light edges, suggesting that large core rotation can occasionally destabilize the plasma, and a low core rotation does not always strongly destabilize. Finally, we see that a high edge rotation has an opposing, destabilizing effect, which indicates the importance of rotation gradients to stability.

While most profile features shown in Fig. 6 show a clear pattern in color, it is important to highlight those with a larger scatter in color, particularly n_e regions and some p regions. This may be a result of low measurement accuracy or strong correlations with other features affecting Shapley calculations. All the input profiles are, of course, highly correlated in a tokamak plasma, with the strongest correlation here being pressure, which is the product of temperature and density. Strongly correlated inputs should not significantly affect ML model performance, but does make the Shapley analysis more difficult, so it is important to remove such features before analysis in future applications.

Focusing on the specific profile features of T_e and T_i in Fig. 7, we see a clear difference in the pattern between the two profiles. The TM risk due to the core T_e rises rapidly, while the trend for core T_i shows little change between 2 and 3.5 keV and an exponential rise at higher values. An exponential fit is shown in Fig. 7 to guide the eye, and shows a larger exponential growth rate for the core T_i at 0.82 keV^{-1} compared to 0.58 keV^{-1} for core T_e. Finally, the off-axis region displays opposite effects between T_e and T_i, notably the off-axis T_i in panel (**e**), strongly stabilizing TMs as its magnitude increases. This shows an underlying difference in the mechanism linking T_e and T_i to the TM onset in this scenario and suggests that broader, flatter temperature profiles are more beneficial for TM stability, especially in the T_i profile, which may be the result of NBI-induced fast ion stabilization.[47]

These insights can be used to inform future elevated q_{min} experiments to minimize TM risk without sacrificing performance. For example, the differences between T_e and T_i in the TM risk suggest that the ratio between ECH (electron heating) and NBI (ion heating) fraction should be tuned to avoid the exponential rise in T_i-induced TM risk. A flatter T_i profile is also found to stabilize TMs, so neutral beams aimed off-axis may result in better passive TM stability than typical on-axis NBI injection. This analysis and its conclusions are presently limited to the elevated q_{min} scenario, but the model interpretation method is general and applicable to any DIII-D regime.

IV. CONCLUSION

Using Shapley analysis, we were able to analyze the plasma profiles to understand their effects on TM stability predictions through a deep learning ML model. This was enabled by an accurate, long time

horizon TM predictor model that was developed for DIII-D and demonstrated in an experiment to accurately predict TMs in real-time. This analysis agrees with generally understood physics observations and theoretical models, such as higher T_e destabilizing TMs while higher rotation stabilizes them. The analysis also led to new TM stability observations, such as the TM risk increasing more rapidly with core T_i compared to core T_e, which suggests high T_e/T_i fraction plasmas may be more stable to TMs. This analysis was performed on an elevated q_{min} experiment and therefore only affects the physics of that scenario, but a larger database study of different scenarios and machines may help uncover the key factors in developing tearing-free plasmas.

We presented two TM prediction models: a "large" model used in a real-time DIII-D TM avoidance experiment, and a "reduced" model used for our profile analysis. The "large" model is best suited for DIII-D experiments as it has the best predictive ability, and is demonstrated in a successful experiment to maintain a TM-free high-performance scenario. Another "reduced" model is developed using simpler full-profile inputs and is used in our Shapley analysis.

Our Shapley analysis framework is also well-suited for analyzing various scenarios in DIII-D. By selecting an appropriate reference distribution, we can tailor the analysis to focus on specific plasma categories, such as advanced non-inductive plasmas, ITER baseline scenario plasmas, or other relevant scenarios. This filtering simplifies the interpretation of the underlying physics because driving TM factors are known to be scenario-dependent.

Many experimental fusion phenomena are challenging to explain using physics-based models, making ML models an attractive alternative due to their high prediction accuracy. With the growing application of ML in fusion experiments, such as Alfvén eigenmodes, ELMs, and general disruptions, it has become more important than ever to understand how these models arrive at their predictions. By applying the Shapley analysis framework introduced in this paper, we can uncover the underlying physics and identify key features that should be controlled to prevent these instabilities, thereby improving the reliability and safety of fusion devices.

Future work includes improving this framework to account for physics effects not explicitly represented by the plasma profiles, such as fast ion population, profile gradients, or the plasma shape. A large input space, however, makes interpretation more complex and correlated, so careful consideration must be made in choosing important parameters. This work is limited in scope to Ip flat-top elevated-qmin shots on DIII-D, so an important extension is a wider database analysis to draw broader and possibly more extrapolable conclusions. In addition, as this framework is general, we believe important insights can be obtained in ELM suppression experiments, L to H transitions and other plasma state transitions, and any complex physics phenomena observed in tokamak plasmas. Finally, important future work would be making this tool easily accessible for between-shot analysis during experiments for scenario development.

ACKNOWLEDGMENTS

This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Fusion Energy Sciences, using the DIII-D National Fusion Facility, a DOE Office of Science user facility, under Award No. DE-FC02-04ER54698. Additionally, this material was supported by the U.S. Department of Energy, under Award No. DE-SC0015480.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Hiro J. Farre-Kaga and Andrew Rothstein contributed equally to this paper.

Hiro J. Farre-Kaga: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (lead). **Andrew Rothstein:** Conceptualization (lead); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Writing – original draft (lead); Writing – review & editing (lead). **Rohit Sonker:** Methodology (equal); Writing – review & editing (equal). **SangKyeun Kim:** Conceptualization (equal); Formal analysis (equal); Writing – review & editing (equal). **Ricardo Shousha:** Data curation (equal); Formal analysis (equal); Writing – review & editing (equal). **Minseok Kim:** Data curation (equal); Writing – review & editing (equal). **Keith Erickson:** Software (lead). **Jeff Schneider:** Supervision (supporting). **Egemen Kolemen:** Conceptualization (equal); Funding acquisition (equal); Resources (equal); Supervision (equal); Visualization (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

APPENDIX A: DETAILED CONSIDERATIONS ON DATABASE PROCESSING AND TM LABELING

- I_P rampup and rampdown data are excluded, as we wished to predict flattop TMs, which can be controlled by changing ECH deposition. We wouldn't want to change ECH deposition during rampup as it may affect the scenario.
- A TM was considered to have occurred if the amplitude of the root mean square (RMS) of the n=1 magnetic fluctuations, called n1rms, signal peaked above 12 G for a continuous 50 ms, along with additional constraints on H_{98} and q_{95} to only consider H-mode plasmas. The onset time of the TM was determined to be when the n1rms first reached 10% of the peak n1rms signal. This means higher m number modes were also included, although they are less likely than the intended 2/1 modes. It also means that n = 2, 3 modes are ignored.
- The n = 1, 2, 3 rms signals, the RMS magnetic signal for n = 1, 2, 3 modes, are excluded from the training set to avoid the model overly relying on these signals, as they are used for labeling.
- The data were chosen to be every 20 ms as this is enough time for diagnostics and EFITs to yield updated results, faster than τ_R and τ_E ensuring the profiles are equilibrated, but not too fast that the model overfits to noise. Actuation, such as modifying ECH deposition and NBI power adjustment, will also affect the plasma on the order of 100 ms, so this is a good compromise.
- The NBI power and torque signals were smoothed to remove the modulation spikes, which are too fast to affect the plasma equilibrium.

- ECH mirror angles are not included as an input. The model is designed to observe physics quantities, such as q profile changes, rather than actuation quantities. This would ensure the models weren't biased to predicting ECCD-suppressed TMs, but it should be a learned effect on the profiles. DIII-D also has a limited variety in ECH mirror angles, making it difficult for models to learn relations between stability and mirror angle.
- We apply a real-time compatible low-pass filter to the inputs and outputs to ensure noise spikes don't cause false positives.
- All the inputs are signals available in real-time in DIII-D, and could be available in most tokamaks. This is to ensure that such a model is applicable to real-time control in present and future reactors.
- The shot time is not provided to the model to avoid overfitting to specific times. Certain scenarios frequently have TMs at specific times, which may cause the model to overfit to time and not learn from the profiles.

APPENDIX B: EVENT CHARACTERIZATION DETAILS

Figure 8 depicts an example TM prediction result to explain the event labeling definitions used for our database. This is a true positive because the threshold is crossed and a TM occurs; however, we define the Δt_{warn} as the *last* time the threshold is crossed. Considering the *first* time, it is crossed results in inflated warning time statistics.

If the last time the threshold is crossed comes after Δt_{lim}, this is considered a false negative even if the event is correctly predicted, since it is too late to actuate on the TM. This paper uses $\Delta t_{lim} = 100$ ms as this approximates the time needed for ECH and most actuation to affect the plasma. Only 21/1476 shots with TMs were in this category because of our high warning times.

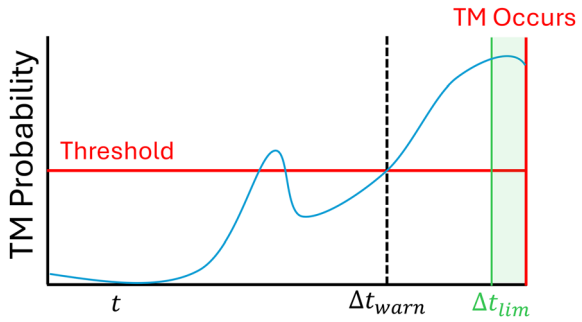

FIG. 8. Diagram of an example TM prediction. Δt_{warn} is the warning time for a TM and Δt_{lim} is the time considered too late to be considered a correct prediction.

APPENDIX C: SHAPLEY TOY MODEL

We consider a toy model to determine the key factors affecting a football team's win percentage

$$\text{winRate} = \frac{1}{\text{norm}}[2 \cdot (\text{Cost}_{\text{squad}})^2 - 20 \cdot (\text{Num}_{\text{injury}}) - 10 \cdot (\text{Age}_{\text{avg}} - 24)^2].$$

With this formula, we can exactly see how each term, the average cost $\text{Cost}_{\text{squad}}$, age Age_{avg}, and injury count $\text{Num}_{\text{injury}}$, contributes to the calculated win rate. However, for our toy model Shapley analysis, this function is hidden, and we consider it as a black-box model that takes input $(\text{Cost}_{\text{squad}}, \text{Age}_{\text{avg}}, \text{Num}_{\text{injury}})$ and outputs the team's win rate.

Figure 9 shows the Shapley analysis of the win rate model using two background distributions (professional with higher costs vs amateur) and two professional teams, detailed in Table II. In (a), the higher squad cost in team 1 gives it a larger Shapley value. The higher value of squad cost (25 M) leads to a 0.1% improvement in the win rate prediction relative to other professionals, which intuitively makes sense as the cost of the squad has a quadratic dependence on the win rate.

Next, we see the importance of the reference distribution when comparing Figs. 9(a) and 9(b). When compared to other professionals, amateur age and injuries will have a meaningful effect on a team's win rate. However, when compared to amateurs that have a significantly lower cost, the professional team's superior cost dominates all other factors. Note that the total Shapley value in (b) is larger than in (a) because Shapley values are relative to the distribution average, which is significantly lower in the amateur group.

APPENDIX D: ERROR ANALYSIS OF SHAPLEY VALUES

The error of individual TM impact Shapley values is evaluated using a Bootstrap sampling method on the background distribution with 100 bootstrap samples. This consists of running Shapley analysis of a given shot 100 times, each with a background distribution that is resampled with replacement from the original dataset. Figures 10 and 11 show the Shapley value and its standard error for each profile in time for shots 199 597 and 199 606. While this

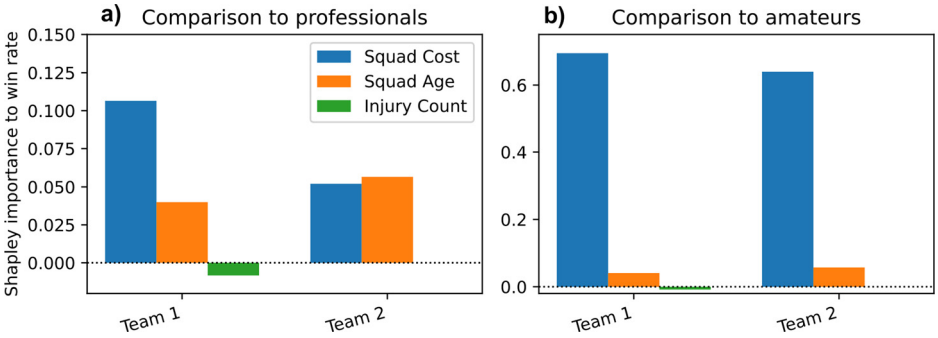

FIG. 9. Shapley values for our toy model. (a) Shapley values between two professional teams show meaningful contributions from cost, age, and injury count, where the difference in cost is the driving factor. (b) Compared to a background of Amateur teams, the cost is now the sole driving factor for the team's win rates and age and injury counts are negligible.

TABLE II. Upper table: Example teams 1 and 2 average cost, age, injury counts, and the corresponding win rate. Lower table: Professional and amateur background distributions used to calculate Shapley values, where the only difference is the range of squad costs, as professionals cost much more than amateurs.

Team	Team 1	Team 2
Cost ($\times 10^6$)	\$25	\$24
Age (years)	22	23
Injury count	8	5
Win rate	0.65	0.62
Context	**Professional**	**Amateur**
Cost ($\times 10^6$)	\$15–\$30	\$13–\$15
Age (years)	20–30	20–30
Injury count	0–10	0–10

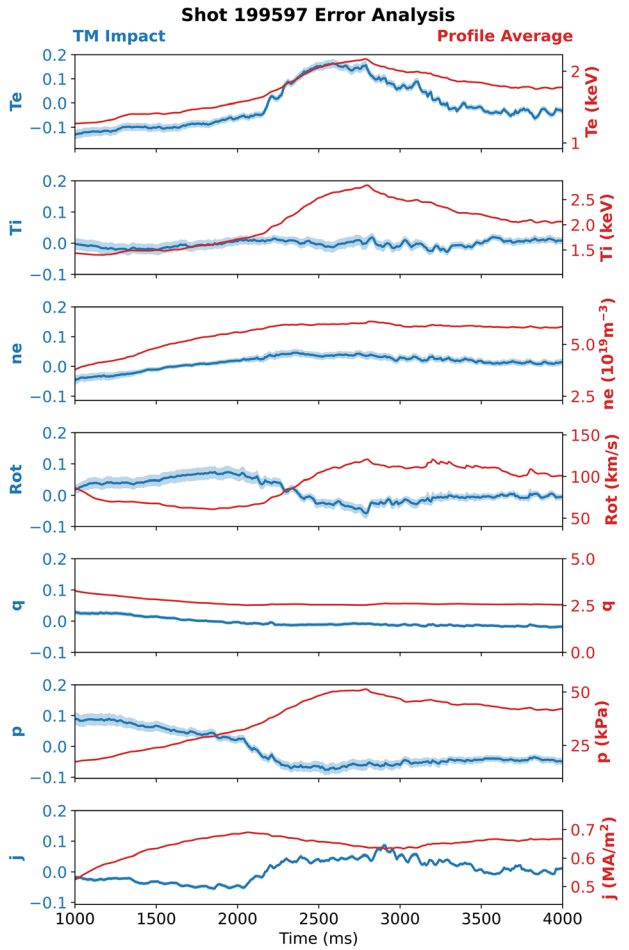

FIG. 10. Error analysis of the total TM impact values for each profile as a function of time in shot 199 597. The blue axis is the total TM impact for each profile, and the red axis is the average profile value. The blue errorbars represent the standard error of the TM impact through Bootstrap sampling. The red shaded region shows the times when a TM was present.

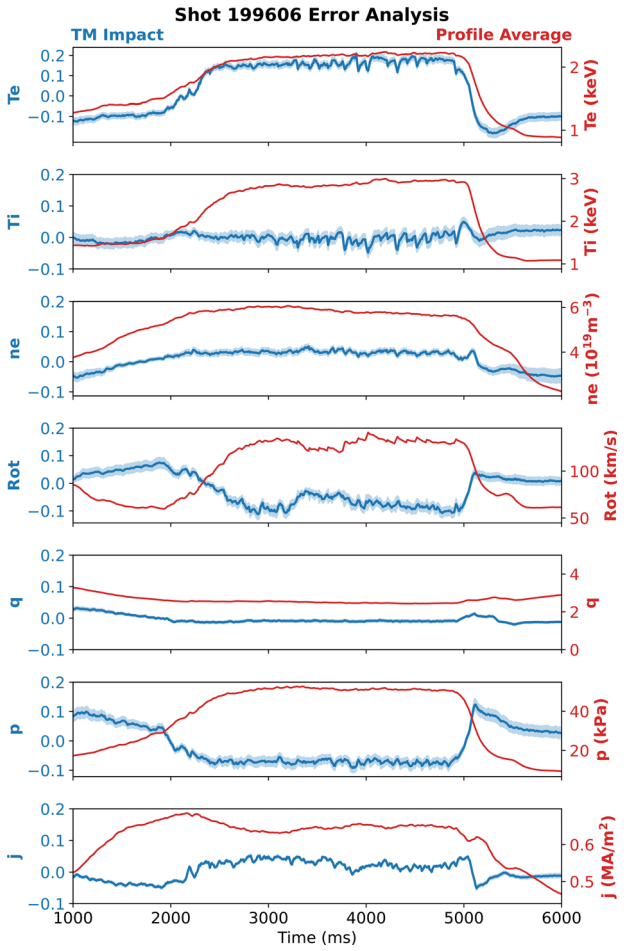

FIG. 11. Error analysis of the total TM impact values for each profile as a function of time in shot 199 606. The blue axis is the total TM impact for each profile, and the red axis is the average profile value. The blue errorbars represent the standard error of the TM impact through Bootstrap sampling. The red shaded region shows the times when a TM was present.

analysis provides the uncertainty for each spatial point in the profile, the sum of all Shapley values and uncertainties is plotted instead, so the temporal evolution can be displayed. While the uncertainties in Figs. 10 and 11 are small, certain regions and profiles have larger uncertainties, such as the rotation before reaching $t = 2000$ ms and the ion temperature in shot 199 606.

An alternate method for visualizing the uncertainty for a wider dataset is by plotting scatter plots, as in Fig. 7. The scatter plots, such as Fig. 7(a), show a clear pattern, but the spread increases at higher Core T_e, suggesting a higher uncertainty and a more difficult regime to draw strong conclusions on TM impact.

APPENDIX E: PROFILE GRADIENT EFFECTS ON TM IMPACT

Profile gradients such as rotation shear and the pressure gradient producing a bootstrap current are important parameters for

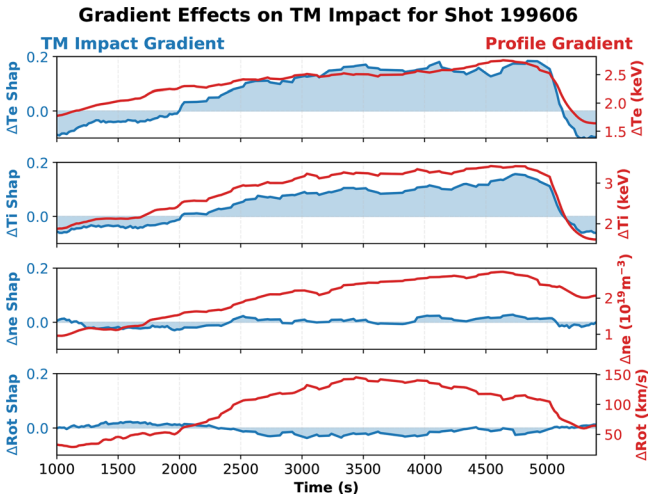

FIG. 12. Analysis of the TM impact from profile gradients as a function of time in shot 199 606. The blue axis reflects the difference in Shapley value between the core and edge of each profile. The red axis is the difference in magnitude between the core and edge for each profile. While these are not direct Shapley values of the gradient, they show how a change in profile gradient affects the TM impact of the profile.

determining TM stability. The TM prediction model uses 1D profiles as inputs and is able to evaluate the importance of gradients through its hidden layers. However, the Shapley analysis presented cannot directly access these effects because gradients are not an explicit input to the model.

One solution is to train a model with both profiles and profile gradients as inputs. This would allow the Shapley analysis to separate the effects of input magnitude and gradients. However, this significantly increases with the input size and its correlation, making it difficult to disentangle different effects. It would also only explain the impact of local gradients and would miss gradients between different regions, which can be more important.[27]

In this section, we provide a simple but informative gradient analysis of the Shapley values presented in Sec. III C. In Fig. 12, we plot the difference between the core and edge profile values, as well as the difference of the Shapley values of those regions, to visualize the correlation between gradients and Shapley values. While this method does not explicitly allocate a Shapley value to the profile gradient, we see the importance of certain profile gradients for the TM model.

As expected from Figs. 4 and 5, as the T_e and T_i gradient increases, the TM impact rises significantly, with the two values being highly correlated. We also see little to no effect of density gradient on the TM impact, as we observed in Sec. III C and in theoretical studies.[46] The density gradient may be slightly destabilizing, but it is small compared to the importance of T_e and T_i gradients.

The rotation gradient plot shows that an increased gradient is more stabilizing, but it is not a significant feature. While in Sec. III C, we saw the impact of high rotation in TM stability; these results suggest rotation shear is not important in this scenario.

REFERENCES

[1]S. M. Morosohk, M. D. Boyer, and E. Schuster, "Accelerated version of NUBEAM capabilities in DIII-D using neural networks," Fusion Eng. Des. **163**, 112125 (2021).

[2]M. Boyer, S. Kaye, and K. Erickson, "Real-time capable modeling of neutral beam injection on NSTX-U using neural networks," Nucl. Fusion **59**, 056008 (2019).

[3]S. Morosohk, A. Pajares, T. Rafiq, and E. Schuster, "Neural network model of the multi-mode anomalous transport module for accelerated transport simulations," Nucl. Fusion **61**, 106040 (2021).

[4]A. Rothstein, A. Jalalvand, J. Abbate, K. Erickson, and E. Kolemen, "Initial testing of Alfvén eigenmode feedback control with machine-learning observers on DIII-D," Nucl. Fusion **64**, 096020 (2024).

[5]Z. Keith, C. Nagpal, C. Rea, and R. A. Tinguely, "Risk-aware framework development for disruption prediction: Alcator C-Mod and DIII-D survival analysis," J. Fusion Energy **43**, 21 (2024).

[6]J. Seo, R. Conlin, A. Rothstein, S. Kim, J. Abbate, A. Jalalvand, and E. Kolemen, "Multimodal Prediction of Tearing Instabilities in a Tokamak," in *2023 International Joint Conference on Neural Networks (IJCNN)* (IEEE, Gold Coast, Australia, 2023), pp. 1–8.

[7]X. K. Ai, W. Zheng, M. Zhang, Y. H. Ding, D. L. Chen, Z. Y. Chen, C. S. Shen, B. H. Guo, N. C. Wang, Z. J. Yang, Z. P. Chen, Y. Pan, B. Shen, B. J. Xiao, and J.-T. Team, "Cross-tokamak deployment study of plasma disruption predictors based on convolutional autoencoder," Plasma Phys. Controlled Fusion **66**, 085015 (2024).

[8]C. Rea, K. Montes, K. Erickson, R. Granetz, and R. Tinguely, "A real-time machine learning-based disruption predictor in DIII-D," Nucl. Fusion **59**, 096016 (2019).

[9]A. Jalalvand, A. A. Kaptanoglu, A. V. Garcia, A. O. Nelson, J. Abbate, M. E. Austin, G. Verdoolaege, S. L. Brunton, W. W. Heidbrink, and E. Kolemen, "Alfvén eigenmode classification based on ECE diagnostics at DIII-D using deep recurrent neural networks," Nucl. Fusion **62**, 026007 (2022).

[10]R. M. Churchill, B. Tobias, Y. Zhu, and DIII-D team, "Deep convolutional neural networks for multi-scale time-series classification and application to tokamak disruption prediction using raw, high temporal resolution diagnostic data," Phys. Plasmas **27**, 062510 (2020).

[11]J. Seo, S. Kim, A. Jalalvand, R. Conlin, A. Rothstein, J. Abbate, K. Erickson, J. Wai, R. Shousha, and E. Kolemen, "Avoiding fusion plasma tearing instability with deep reinforcement learning," Nature **626**, 746–751 (2024).

[12]J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. De Las Casas, C. Donner, L. Fritz, C. Galperti, A. Huber, J. Keeling, M. Tsimpoukelli, J. Kay, A. Merle, J.-M. Moret, S. Noury, F. Pesamosca, D. Pfau, O. Sauter, C. Sommariva, S. Coda, B. Duval, A. Fasoli, P. Kohli, K. Kavukcuoglu, D. Hassabis, and M. Riedmiller, "Magnetic control of tokamak plasmas through deep reinforcement learning," Nature **602**, 414–419 (2022).

[13]Z. Liu and F. Xu, "Interpretable neural networks: Principles and applications," Front. Artif. Intell. **6**, 974295 (2023).

[14]S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Curran Associates, Inc., Red Hook, NY, USA, 2017), pp. 4768–4777.

[15]K. E. J. Olofsson, D. A. Humphreys, and R. J. L. Haye, "Event hazard function learning and survival analysis for tearing mode onset characterization," Plasma Phys. Controlled Fusion **60**, 084002 (2018).

[16]K. Olofsson, B. Sammuli, and D. Humphreys, "Hazard function exploration of tokamak tearing mode stability boundaries," Fusion Eng. Des. **146**, 1476–1479 (2019).

[17]Y. Fu, D. Eldon, K. Erickson, K. Kleijwegt, L. Lupin-Jimenez, M. D. Boyer, N. Eidietis, N. Barbour, O. Izacard, and E. Kolemen, "Machine learning control for disruption and tearing mode avoidance," Phys. Plasmas **27**, 022501 (2020).

[18]K. Olofsson, C. Akçay, and B. Sammuli, "Database-wide hazard modelling of the onset of DIII-D tearing modes with field features," J. Plasma Phys. **88**, 895880503 (2022).

[19]K. E. J. Olofsson, C. Akcay, X. Sun, B. S. Sammuli, and R. Nazikian, "Large-scale tearing-mode hazard function analysis with standard matched equilibrium reconstructions," Plasma Phys. Controlled Fusion **67**, 065039 (2025).

[20]C. Nagpal, X. Li, and A. Dubrawski, "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks," IEEE J. Biomed. Health Inf. **25**, 3163–3175 (2021).

[21]S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," Nat. Biomed. Eng. **2**, 749–760 (2018).
[22]M. Landreman, J. Y. Choi, C. Alves, P. Balaprakash, R. M. Churchill, R. Conlin, and G. Roberg-Clark, "How does ion temperature gradient turbulence depend on magnetic geometry? Insights from data and machine learning," J. Plasma Phys. **91**, E120 (2025).
[23]T. Pyragius, C. Colgan, H. Lowe, F. Janky, M. Fontana, Y. Cai, and G. Naylor, "Application of interpretable machine learning for cross-diagnostic inference on the ST40 spherical tokamak," arXiv:2407.18741 (2024).
[24]L. Bardóczi, N. Richner, and N. Logan, "The onset distribution of rotating m, n = 2, 1 tearing modes and its consequences on the stability of high-confinement-mode plasmas in DIII-D," Nucl. Fusion **63**, 126052 (2023).
[25]L. Bardóczi, N. J. Richner, J. Zhu, C. Rea, and N. C. Logan, "Empirical probability and machine learning analysis of m, n = 2, 1 tearing mode onset parameter dependence in DIII-D H-mode scenarios," Phys. Plasmas **30**, 092505 (2023).
[26]F. Turco, T. Luce, W. Solomon, G. Jackson, G. Navratil, and J. Hanson, "The causes of the disruptive tearing instabilities of the ITER Baseline Scenario in DIII-D," Nucl. Fusion **58**, 106043 (2018).
[27]L. Bardoczi, N. Richner, N. Logan, E. Strait, C. Holcomb, J. Zhu, and C. Rea, "The root cause of disruptive NTMs and paths to stable operation in DIII-D ITER baseline scenario plasmas," Nucl. Fusion **64**, 126005 (2024).
[28]N. Richner, L. Bardóczi, J. Callen, R. La Haye, N. Logan, and E. Strait, "Use of differential plasma rotation to prevent disruptive tearing mode onset from 3-wave coupling," Nucl. Fusion **64**, 106036 (2024).
[29]A. S. Glasser, A. H. Glasser, R. Conlin, and E. Kolemen, "An ideal MHD δw stability analysis that bypasses the Newcomb equation," Phys. Plasmas **27**, 022114 (2020).
[30]H. R. Wilson, "Neoclassical tearing modes," Fusion Sci. Technol. **45**, 123 (2004).
[31]R. Carrera, R. D. Hazeltine, and M. Kotschenreuther, "Island bootstrap current modification of the nonlinear dynamics of the tearing mode," Phys. Fluids **29**, 899–902 (1986).
[32]R. Fitzpatrick, "Helical temperature perturbations associated with tearing modes in tokamak plasmas," Phys. Plasmas **2**, 825–838 (1995).
[33]J. W. Connor, F. L. Waelbroeck, and H. R. Wilson, "The role of polarization current in magnetic island evolution," Phys. Plasmas **8**, 2835–2848 (2001).
[34]T. N. Carlstrom, G. L. Campbell, J. C. DeBoo, R. Evanko, J. Evans, C. M. Greenfield, J. Haskovec, C. L. Hsieh, E. McKee, R. T. Snider, R. Stockdale, P. K. Trost, and M. P. Thomas, "Design and operation of the multipulse Thomson scattering diagnostic on DIII-D (invited)," Rev. Sci. Instrum. **63**, 4901–4906 (1992).
[35]R. P. Seraydarian and K. H. Burrell, "Multichordal charge-exchange recombination spectroscopy on the DIII-D tokamak," Rev. Sci. Instrum. **57**, 2012–2014 (1986).
[36]D. Wróblewski, K. H. Burrell, L. L. Lao, P. Politzer, and W. P. West, "Motional Stark effect polarimetry for a current profile diagnostic in DIII-D," Rev. Sci. Instrum. **61**, 3552–3556 (1990).
[37]L. Lao, H. St. John, R. Stambaugh, A. Kellman, and W. Pfeiffer, "Reconstruction of current profile parameters and plasma shapes in tokamaks," Nucl. Fusion **25**, 1611–1622 (1985).
[38]R. Sweeney, W. Choi, M. Austin, M. Brookman, V. Izzo, M. Knolker, R. La Haye, A. Leonard, E. Strait, and F. Volpe, and DIII-D Team. "Relationship between locked modes and thermal quenches in DIII-D," Nucl. Fusion **58**, 056022 (2018).
[39]R. Shousha, J. Seo, K. Erickson, Z. Xing, S. Kim, J. Abbate, and E. Kolemen, "Machine learning-based real-time kinetic profile reconstruction in DIII-D," Nucl. Fusion **64**, 026006 (2024).
[40]Z. Xing, D. Eldon, A. Nelson, M. Roelofs, W. Eggert, O. Izacard, A. Glasser, N. Logan, O. Meneghini, S. Smith, R. Nazikian, and E. Kolemen, "CAKE: Consistent Automatic Kinetic Equilibrium reconstruction," Fusion Eng. Des. **163**, 112163 (2021).
[41]C. T. Holcomb, J. R. Ferron, T. C. Luce, T. W. Petrie, J. M. Park, F. Turco, M. A. Van Zeeland, M. Okabayashi, C. T. Lasnier, J. M. Hanson, P. A. Politzter, Y. In, A. W. Hyatt, R. J. La Haye, and M. J. Lanctot, "Steady state scenario development with elevated minimum safety factor on DIII-D," Nucl. Fusion **54**, 093009 (2014).
[42]E. Kolemen, A. Welander, R. La Haye, N. Eidietis, D. Humphreys, J. Lohr, V. Noraky, B. Penaflor, R. Prater, and F. Turco, "State-of-the-art neoclassical tearing mode control in DIII-D using real-time steerable electron cyclotron current drive launchers," Nucl. Fusion **54**, 073020 (2014).
[43]T. C. Luce, C. C. Petty, W. H. Meyer, C. T. Holcomb, K. H. Burrell, and L. J. Bergsten, "Method for correction of measured polarization angles from motional Stark effect spectroscopy for the effects of electric fields," Plasma Phys. Controlled Fusion **58**, 125010 (2016).
[44]R. J. La Haye, R. J. Buttery, S. Guenter, G. T. A. Huysmans, M. Maraschek, and H. R. Wilson, "Dimensionless scaling of the critical beta for onset of a neoclassical tearing mode," Phys. Plasmas **7**, 3349–3359 (2000).
[45]R. J. Buttery, S. Gerhardt, A. Isayama, R. J. L. Haye, E. J. Strait, D. P. Brennan, P. Buratti, D. Chandra, S. Coda, J. D. Grassie, P. Gohil, M. Gryaznevich, J. Hobirk, C. Holcomb, D. F. Howell, G. Jackson, M. Maraschek, A. Polevoi, H. Reimerdes, D. Raju, S. Sabbagh, S. Saarelma, M. Schaffer, and A. Sen, "Multimachine extrapolation of neoclassical tearing mode physics to ITER," Proceedings of the 22nd IAEA Fusion Energy Conference, 2008.
[46]Y. Ming and W. Wang, "Influence of plasma density gradient on the tearing mode with the poloidal shear flow," AIP Adv. **14**, 115103 (2024).
[47]H. Cai, "Influence of energetic ions on neoclassical tearing modes," Nucl. Fusion **56**, 126016 (2016).