Interpreting AI for Fusion: an application to Plasma Profile Analysis for Tearing Mode Stability

Hiro J Farre-Kaga,[1,2] 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]
[1)] *Princeton University, Princeton, NJ, USA*
[2)] *Princeton Plasma Physics Laboratory Princeton, NJ, USA*
[3)] *Carnegie Mellon University, Pittsburgh, PA, USA*

AI models have demonstrated strong predictive capabilities for various tokamak instabilities—including tearing modes (TM), ELMs, 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 first 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, peaked rotation profiles are lightly stabilizing, but core electron temperature and density profile shape play the primary role in TM stability. This work offers a generalizable ML-based event prediction methodology, from training to physics-driven interpretability, bridging the gap between physics understanding and opaque ML models.

Keywords: Machine Learning, Interpretable AI, Tearing Modes, Tokamak, DIII-D, Plasma Control

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[5–9], and reinforcement learning controllers[10,11]. 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[7], that offer interpretable results at the trade-off of simpler ML architectures with lower accuracies. However simple black-box models such as multilayer perceptrons (MLP) 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[12]

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[13] 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 TM stability and explain what specific profile features, such as rotation peaking, led to the avoidance of TMs. Previous TM prediction and stability models utilized just scalar parameters[14,15] or full plasma profiles with no stability interpretation[6]. We improve on these with a ML-based deep survival machine model[16] 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 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[17], there are minimal applications to fusion and has not been used for understanding fusion experiments[18,19], 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 these instabilities, often with conflicting results or limited to scenario-specific operating regimes such as the DIII-D ITER baseline[20–22]. Other approaches to better understand TM stability involve using a physics code such as STRIDE to calculate the the classical Δ' stability parameter[23], however, this has not been validated in experiments. The results from analyzing our TM prediction model with Shapley analysis 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 Section II. Section III describes the TM predictor model results, followed by interpretation for the TM preemptive avoidance experiment. Then we use Shapley values to draw conclusions more broadly about profile-based TM stability. Finally Section IV summarizes our findings and describes the future work to improve TM models and their interpretation.

II. BACKGROUND

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

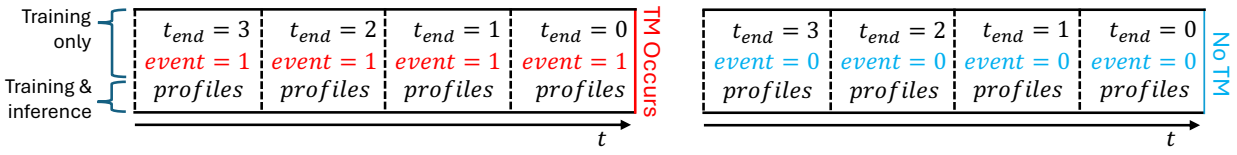

FIG. 1. Depiction of the survival regression training scheme. In training, the model is input t_{end} representing the time until the end of the sequence, *event* representing whether a TM occurs at the end of the sequence (1) or not (0), and $profiles$ representing the set of diagnostics and inputs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

A. Database processing and TM labelling

The model was trained on all DIII-D shots identified to have the required data between shots 140000 to 195000, resulting in 6050 shots of which 1476 contained n=1 TMs, which amounts to 677494 timesteps each with 42 parameters. The data needed was Thomson Scattering, Charge Exchange Recombination Spectroscopy, Motional Stark Effect, magnetics for EFIT reconstructions 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. Additionally, the planned control actions of applying ECCD 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.

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

- I_P rampup and rampdown are excluded as we are only targeting TMs in I_P flattop.
- A TM was considered to have occurred if the n1rms signal peaked above 12G for a continuous 50ms 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.
- Magnetic fluctuation signals like the n1,2,3rms signals for 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 is taken every 20ms as this is enough time for diagnostics and EFITs to yield updated results, faster than τ_R and τ_E ensuring the profile are equilibrated, but not too fast that the model overfits to noise. Actuation such as mirror steering and NBI power adjustment will also affect the plasma on the order of $100ms$, 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[16]. The framework allows for easy-to-use event prediction, and has been proven 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 is representative of the underlying physics. Fig 1 depicts the training scheme, where in training we input the plasma parameters such as the profiles, the time until the end of 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 are not known. This training setup informs the model whether the given plasma parameters will be unstable t timesteps in the future.

Survival regression is a statistical scheme that provides a probability of an event occurring at any time within a user-chosen time horizon, $t_{horizon}$, given a set of input features. 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 as the 'symptoms'. This is therefore applicable to plasma disruptions in general, and to specific MHD events such as the onset of a 2/1 TM.

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[13] 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

Input	Source
Electron Temperature profile (Te)	RTCAKENN
Electron density profile (ne)	RTCAKENN
Ion temperature profile (Ti)	RTCAKENN
Ion rotation profile (Ω)	RTCAKENN
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

TABLE I. Input parameters used in the analysis and their corresponding sources. EFITRT2 is real-time magnetic equilibrium reconstruction using magnetic diagnostics and motional stark effect. All profiles from RTCAKENN are reduced to the 4 main PCA components.

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. To demonstrate this technique, we have provided an illustrative toy example in Appendix C.

In our TM study, we use the 11 shots from our dedicated DIII-D experiment as the background distribution. 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 Section III A. The DIII-D experiment on preemptive TM avoidance is explained in detail in Section III B, showing successful

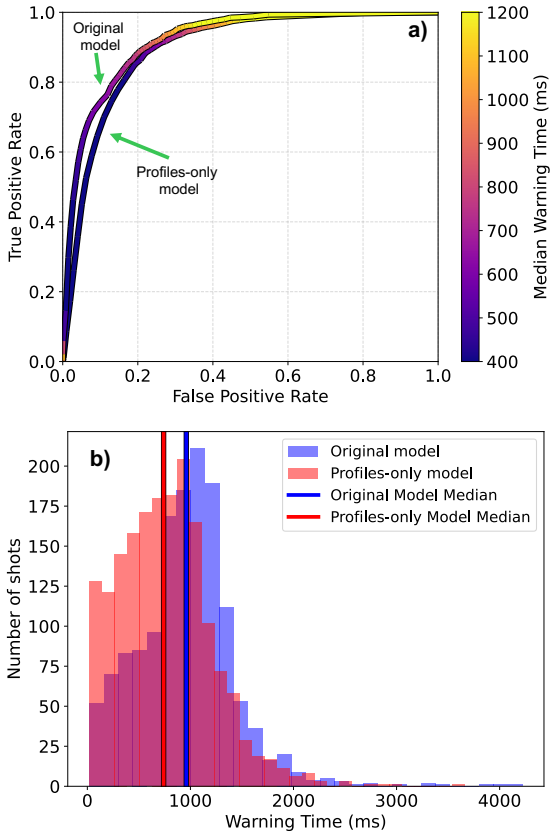

FIG. 2. a) ROC curve for the RTCAKENN-based TM predictor used in the experiment. The original model had an AUROC score of 0.92, compared to 0.89 for the profiles-only model. b) Warning times histogram for a typical threshold of 0.2 shows the majority of TMs are predicted over 500ms in advance, allowing for flexibility in actuation.

TM avoidance and detailing the model's results shot-by-shot. We look into specific time-slice profiles and discuss the interpretation of the model using Shapley analysis in Section III C, where the effects of ECCD on equilibrium profiles are shown to stabilize TMs. Finally in Section 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 was trained to predict DIII-D n=1 TMs using the parameters shown in Table I, including the real-time kinetic profiles RTCAKENN[24], an ML surrogate model for CAKE[25], as well as external heating and actuation, and EFITRT2 scalars. The decisions on database selection and processing are listed in II A. Most decisions were made to fit the experimental design, which required a real-time capable flattop TM predictor with enough warning time to steer mirrors and affect the equilibrium (around 200ms).

The basic performance metrics are shown in Fig. 2, demonstrating high performance metrics with warning times around

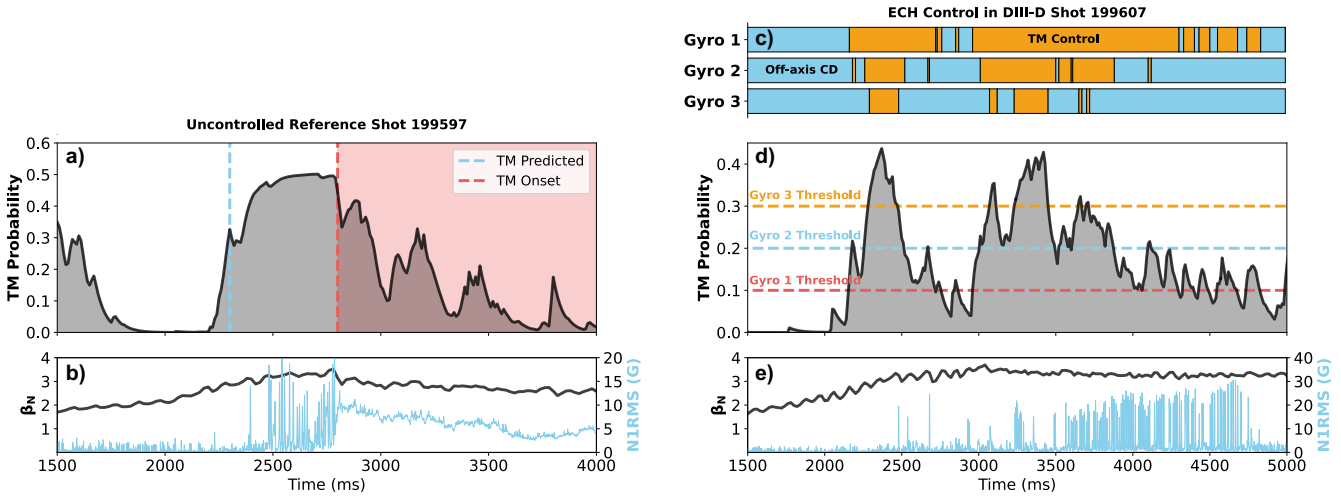

FIG. 3. Demonstration of active preemptive TM suppression via ECH steering. a),b) shows the uncontrolled reference shot, which resulted in a TM at t=2700ms, predicted at t=2300ms. c), d), e) shows a shot with TM suppression via ECH steering. In c), blue represents gyrotrons aimed at an off-axis location and orange represents gyrotrons aimed at the q=2 surface for TM control.

1000ms an AUROC score of 0.92. This allows for flexibility in actuation to avoid the oncoming TM, such as steering the electron cyclotron heating mirrors in the case of our experiment. Further notes on the definition of warning times and classification details are given in Appendix B.

The model architecture was a Deep Survival Machine withh specific parameters of: an MLP with layers of dimension 42×100, 100×1000, a log-normal distribution, batch size of 1000, learning rate of 1×10^{-5}, $k = 3$ survival distributions, and 10000 epochs.

A second model whose inputs are only the 7 RTCAKENN profiles was also trained but not used in experiment. This model is used in section III C to study how profile changes affect the TM risk without confounding scalar inputs. Specifically, β_N, input heating and some shape parameters were found to have consistent, large Shapley values which made analysis of profile importance more difficult. This model of course has worse performance metrics as it used fewer diagnostics and inputs, but the AUROC score of 0.89 is not significantly worse. An important change is in the warning time distribution, which can be seen in fig. 2b).

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. This publication[26] by the authors of this paper describes the experiment and its results in detail.

Previous experiments have shown the potential of preemptive suppression of TMs using electron cyclotron current drive[27] by steering 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, where we predict TMs before they appear, and steer the mirrors to stabilize the profiles, enabling fully tearing free operation. The experiment successfully demonstrated the suppression scheme as can be seen in Fig. 3, resulting in the tearing free operation of previously unstable conditions.

More 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 ECH steering. The experiment started at shot 199597 which was a reference elevated q_{min} shot, and ended at 199607. The reference was well predicted, as shown in Fig. 3, followed by another correct unstable prediction in 199598 and 199599. Shots 199605, 199606, 199607 were the same unstable conditions with predicted TMs, but the active steering of ech successfully avoided TMs. Shots 199600 and 199601 used additional gyrotrons which led to passively stable conditions, correctly predicted again. The only failure of the predictive model was shots 199602 and 199603 which were run at lower plasma current, creating n=3 modes that triggered n=1 TMs. This resulted in a 82% success rate, where the 2 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 199597 as the no-control baseline, where a TM occurred and was predicted 500ms in advance; and shot 199607, which had the same actuation and

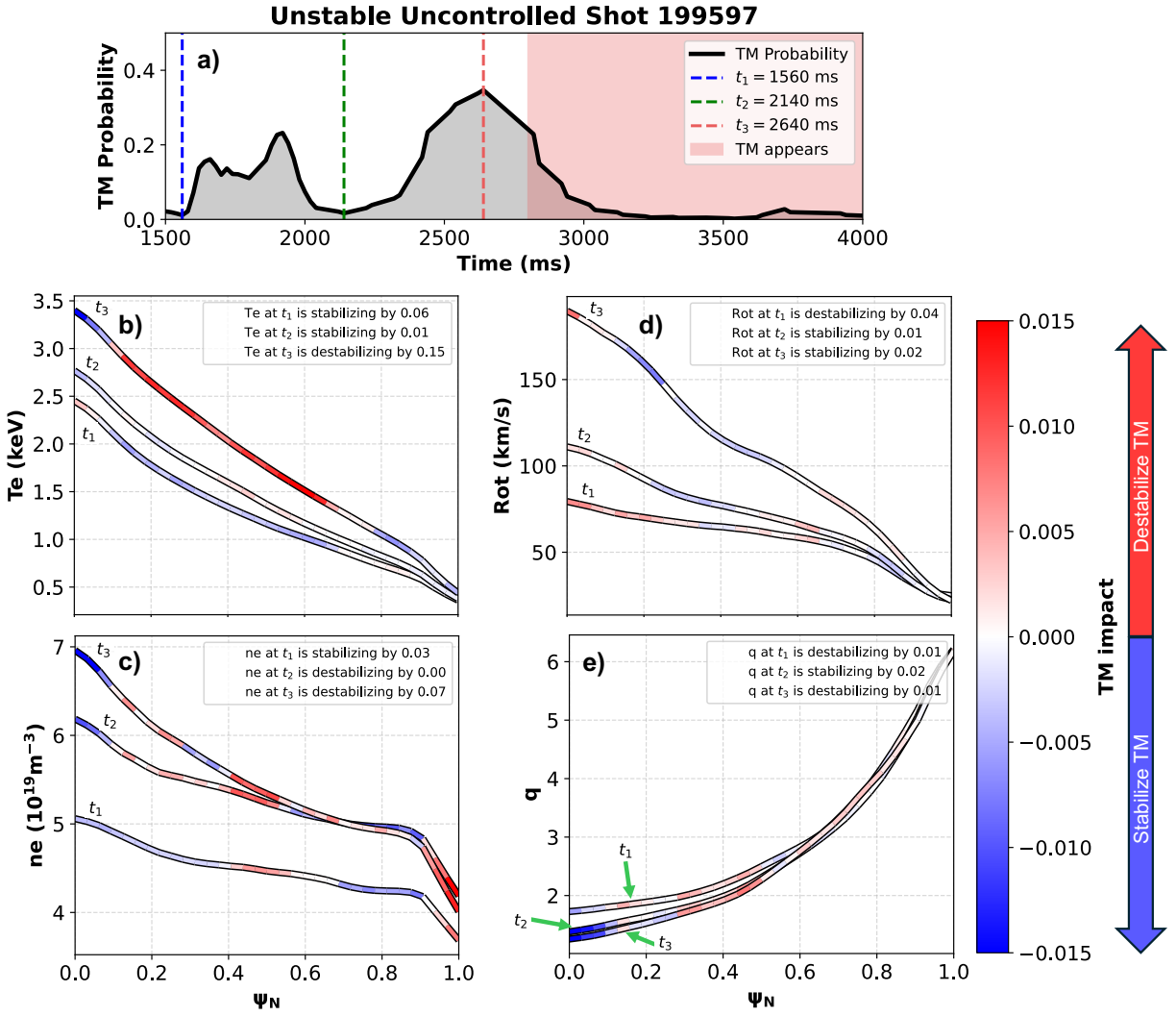

FIG. 4. Shapley analysis for the unstable uncontrolled reference shot 199597. 3 timesteps are chosen, $t_1 = 1560ms$ at the stable start of the shot, $t_2 = 2140ms$ right before the TM is predicted, and $t_3 = 2640ms$ where tearing risk is maximal, and a TM is about to occur at $t = 2700ms$. The color of the profiles represents the tearing impact, where red is destabilizing (positive TM impact), blue is stabilizing (negative TM impact), and white has no impact. The total impact on stability of the profile, or the sum of Shapley values, is in the legend.

conditions as the reference, with the sole difference that gyrotrons were steered from an off-axis current drive position to the q=2 surface whenever TMs were predicted.

For the following analysis, we use the new profile-only model whose inputs were solely the 7 profiles, which explains the small differences in TM probability predictions for the same shots between Fig. 3 and Fig. 4, 5. 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, which is correlated to the pressure profile, to avoid ambiguity in Shapley values. Using only profiles allows a study of their true impact on TM stability.

For this analysis, we refer to the calculated "Shapley value" as "TM impact" to make the interpretation of the values explicitly clear. In Fig. 4,5 positive TM impacts (positive Shapley value) causes TM destabilization and is represented by redder colors while negative TM impacts (negative Shapley value) causes TM stabilization and is represented by bluer colors.

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 small 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.

Similarly, the evolution of TM impact values for the rotation profile provides an 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 per-

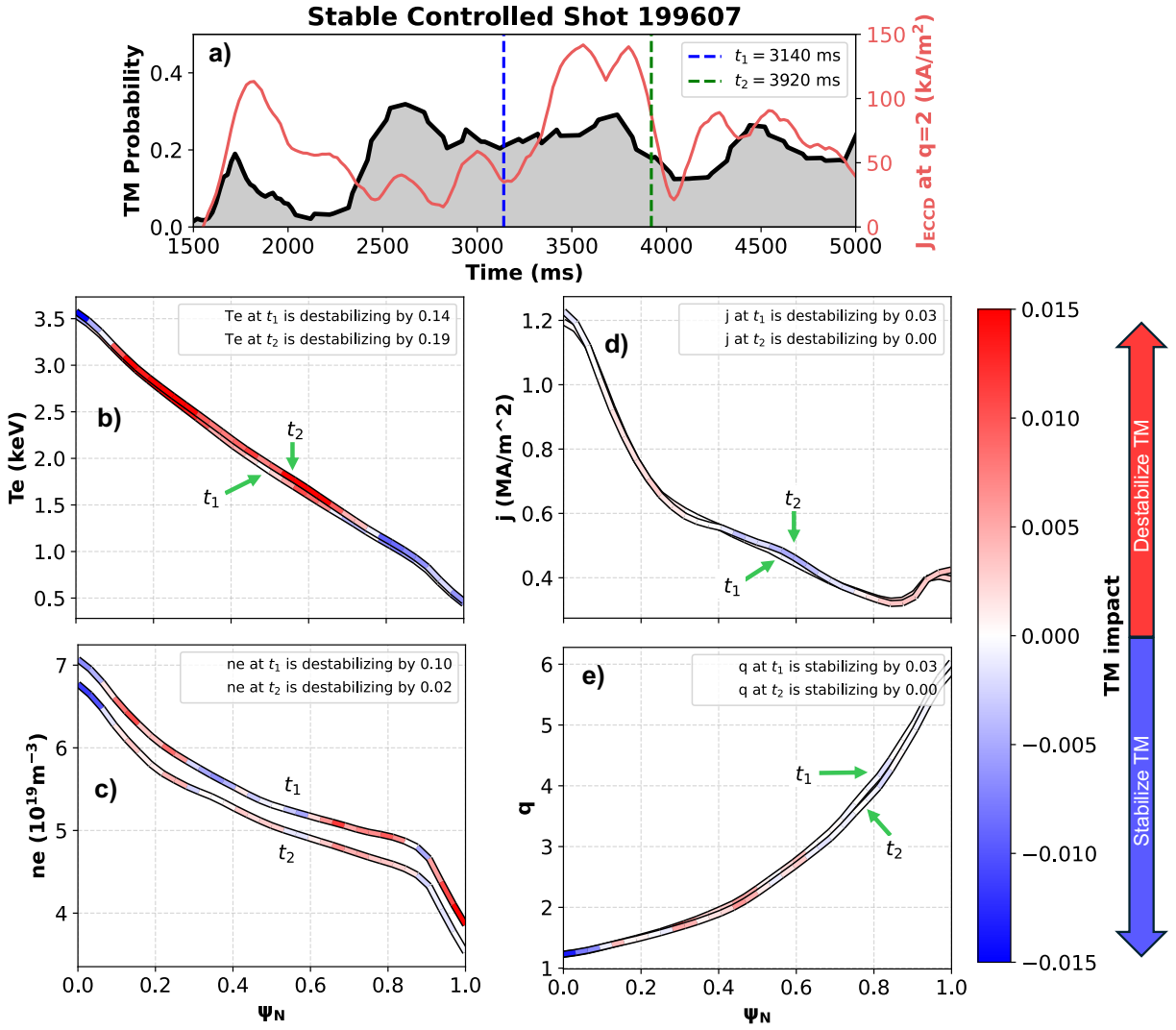

FIG. 5. Shapley analysis for the ECH controlled stable shot 199607. Two timesteps are chosen, $t_1 = 3140ms$ when J_{ECCD} at $q = 2$ is low and the tearing risk is high, and $t_2 = 3920ms$ where gyrotrons have been steered so J_{ECCD} at $q = 2$ is high and the tearing risk begins to drop. The rotation profile is not shown, as the small difference between t_1 and t_2 had little impact.

spective. The higher edge rotation has the opposite destabilizing effect, which suggests that rotation gradient or peaking is an important feature. However, 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 0.06, which is significantly lower than the the 0.21 destabilizing impact of the T_e increase, suggesting that this TM is primarily caused by high T_e.

The density profile evolution also has a relatively low impact to TM probability, but has interesting features. The peaked density profile at t_3 is more destabilizing than t_2 overall, with an important feature being the higher pedestal at t_2. 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_2 and t_3, 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 gyrotron steering 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 gyrotron steering 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 gyrotrons have been aimed away from the $q = 2$ surface, and t_2, where the gyrotrons have been aimed at the $q = 2$ surface for 800ms and the TM probability has begun to drop.

Several profile effects are expected from the gyrotrons steering to the $q = 2$ surface. Primarily, we expect increased electron temperature heating to the region, as seen in Fig. 5b). Since the gyrotron beams were set up to drive current, 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. 5d), as is intended from driving ECCD at the $q = 2$ surface where 2/1 TMs appear. Finally, the lower density at the pedestal region in Fig. 5c) has an overall stabilizing effect, although the density profile TM impact colors do not have a simple pattern as the other profiles do. The two q profiles in Fig. 5e) are too similar to draw any conclusions on time-dependence, but the key features seen for the reference shot still remain, such as the stabilizing q_{min} value and the destabilizing $q = 2$ value.

Overall we observe the tree key profile changes from ECH steering, namely localized current drive, electron heating and density pumpout, have an important effect on TM stability, suggesting that ECH steering 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 for a wider database analysis, providing insight on 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 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 a higher pedestal temperatures contributing to lower TM risk.

The profiles largely derived from magnetic measurements, j and q, have generally smaller TM impacts. This may be a result of inaccurate current measurements or may be an effect of all shots covered in this analysis having similar j and q profiles. With less variation in j and q profiles we would expect 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 should reduce TM risk.

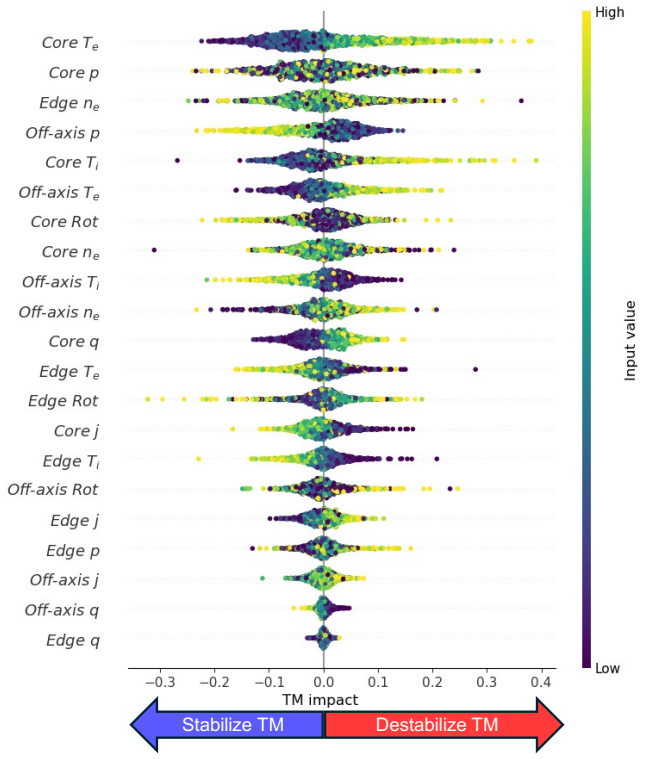

FIG. 6. A general Shapley analysis of all inputs, ordered by influence for TM stability predictions in the dedicated experiment. Each row is a histogram of TM impact for a given value, with its color indicated the magnitude of the value. For example, Core T_e has a wide distribution and thus large mean absolute value of TM impact, suggesting high importance for TM prediction. Additionally, the lighter the colors, or the higher the Core T_e, the more destabilizing it is.

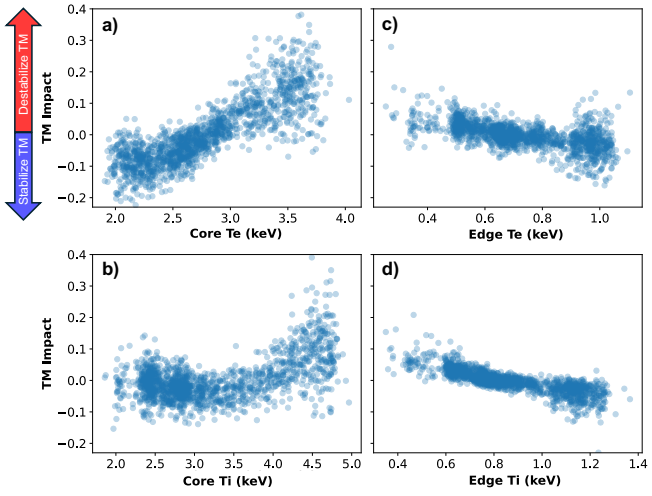

FIG. 7. Average values for T_e and T_i across core and edge regions. Each point represents the average of the profile region at one timeslice during the control experiment. Each point is plotted along the x-axis with the average value for the profile region and the y-value represents the TM impact. From these plots, we can draw conclusions about the correlations between average values in a profile region to its TM impact.

The rotation profile for all regions has a low TM impact for low rotation, but rises (both negatively and positively) as the rotation increases, resulting in a dark center with light edges. This indicates that a high rotation can be both favorable or unfavorable depending on the other plasma parameters. However, in general a high Core rotation with a low edge, or a highly peaked profile, is the most stabilizing shape, suggesting the rotation gradient may be influential.

While most profile features shown in Fig. 6 show a clear pattern in color, it is important to highlight those with a large scatter in color and no clear interpretation, such as Core p and Edge n_e. 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, and makes 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. While the TM risk due to core T_e rises linearly with its magnitude, core T_i shows little change in TM risk between 2keV and 3.5keV and an exponential rise at higher values. Both Edge T_i and T_e show that a higher pedestal is more stable to TMs. This shows an underlying difference in the mechanism linking core T_e and T_i to TM onset in this scenario and suggests that broader, flatter temperature profiles are more beneficial for TM stability.

These insights can be used to inform future experiments to minimize TM risk without sacrificing performance. For example, the differences between T_i and T_e in 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_e profile is also found to stabilize TMs, so a broad ECH heating profile may result in better passive TM stability than strongly localized heating.

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 proven in experiment to accurately predict TMs in real-time. This analysis validated generally understood physics observations 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 exponentially with Core T_i, compared to a linear increase with 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 reflects 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.

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.

ACKNOWLEDGMENTS

This material is 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 DE-FC02-04ER54698. Additionally, this material is supported by the U.S. Department of Energy, under Award DE-SC0015480.

REFERENCES

[1] Shira M. Morosohk, Mark D. Boyer, and Eugenio Schuster. Accelerated version of NUBEAM capabilities in DIII-D using neural networks. *Fusion Engineering and Design*, 163:112125, February 2021.

[2] M.D. Boyer, S. Kaye, and K. Erickson. Real-time capable modeling of neutral beam injection on NSTX-U using neural networks. *Nuclear Fusion*, 59(5):056008, May 2019.

[3] S.M. Morosohk, A. Pajares, T. Rafiq, and E. Schuster. Neural network model of the multi-mode anomalous transport module for accelerated transport simulations. *Nuclear Fusion*, 61(10):106040, October 2021.

[4] Andrew Rothstein, Azarakhsh Jalalvand, Joseph Abbate, Keith Erickson, and Egemen Kolemen. Initial testing of Alfvén eigenmode feedback control with machine-learning observers on DIII-D. *Nuclear Fusion*, 64(9):096020, September 2024.

[5] Zander Keith, Chirag Nagpal, Cristina Rea, and R. Alex Tinguely. Risk-Aware Framework Development for Disruption Prediction: Alcator C-Mod and DIII-D Survival Analysis. *Journal of Fusion Energy*, 43(1):21, June 2024.

[6] Jaemin Seo, Rory Conlin, Andrew Rothstein, SangKyeun Kim, Joseph Abbate, Azarakhsh Jalalvand, and Egemen Kolemen. Multimodal Prediction of Tearing Instabilities in a Tokamak. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Gold Coast, Australia, June 2023. IEEE.

[7] C. Rea, K.J. Montes, K.G. Erickson, R.S. Granetz, and R.A. Tinguely. A real-time machine learning-based disruption predictor in DIII-D. *Nuclear Fusion*, 59(9):096016, September 2019.

[8] Azarakhsh Jalalvand, Alan A. Kaptanoglu, Alvin V. Garcia, Andrew O. Nelson, Joseph Abbate, Max E. Austin, Geert Verdoolaege, Steven L. Brunton, William W. Heidbrink, and Egemen Kolemen. Alfvén eigenmode classification based on ECE diagnostics at DIII-D using deep recurrent neural networks. *Nuclear Fusion*, 62(2):026007, February 2022.

[9] 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. *Physics of Plasmas*, 27(6):062510, June 2020.

[10] Jaemin Seo, SangKyeun Kim, Azarakhsh Jalalvand, Rory Conlin, Andrew Rothstein, Joseph Abbate, Keith Erickson, Josiah Wai, Ricardo Shousha, and Egemen Kolemen. Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature*, 626(8000):746–751, February 2024.

[11] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego De Las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, February 2022.

[12] Zhuoyang Liu and Feng Xu. Interpretable neural networks: principles and applications. *Frontiers in Artificial Intelligence*, 6:974295, October 2023.

[13] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[14] K E J Olofsson, D A Humphreys, and R J La Haye. Event hazard function learning and survival analysis for tearing mode onset characterization. *Plasma Physics and Controlled Fusion*, 60(8):084002, August 2018.

[15] K.E.J. Olofsson, B.S. Sammuli, and D.A. Humphreys. Hazard function exploration of tokamak tearing mode stability boundaries. *Fusion Engineering and Design*, 146:1476–1479, September 2019.

[16] Chirag Nagpal, Xinyu Li, and Artur Dubrawski. *Deep Survival Machines* : Fully Parametric Survival Regression and Representation Learning for Censored Data With Competing Risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, August 2021.

[17] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749–760, October 2018.

[18] Matt Landreman, Jong Youl Choi, Caio Alves, Prasanna Balaprakash, R. Michael Churchill, Rory Conlin, and Gareth Roberg-Clark. How does ion temperature gradient turbulence depend on magnetic geometry? Insights from data and machine learning, February 2025. arXiv:2502.11657 [physics].

[19] Tadas Pyragius, Cary Colgan, Hazel Lowe, Filip Janky, Matteo Fontana, Yichen Cai, and Graham Naylor. Application of interpretable machine learning for cross-diagnostic inference on the ST40 spherical tokamak, July 2024. arXiv:2407.18741 [physics].

[20] F. Turco, T.C. Luce, W. Solomon, G. Jackson, G.A. Navratil, and J.M. Hanson. The causes of the disruptive tearing instabilities of the ITER Baseline Scenario in DIII-D. *Nuclear Fusion*, 58(10):106043, October 2018.

[21] L. Bardoczi, N.J. Richner, N.C. Logan, E.J. Strait, C.T. Holcomb, J. Zhu, and C. Rea. The root cause of disruptive NTMs and paths to stable operation in DIII-D ITER baseline scenario plasmas. *Nuclear Fusion*, 64(12):126005, December 2024.

[22] N.J. Richner, L. Bardóczi, J.D. Callen, R.J. La Haye, N.C. Logan, and E.J. Strait. Use of differential plasma rotation to prevent disruptive tearing mode onset from 3-wave coupling. *Nuclear Fusion*, 64(10):106036, October 2024.

[23] Alexander S. Glasser, A. H. Glasser, Rory Conlin, and Egemen Kolemen. An ideal MHD δW stability analysis that bypasses the Newcomb equation. *Physics of Plasmas*, 27(2):022114, February 2020.

[24] Ricardo Shousha, Jaemin Seo, Keith Erickson, Zichuan Xing, SangKyeun Kim, Joseph Abbate, and Egemen Kolemen. Machine learning-based real-time kinetic profile reconstruction in DIII-D. *Nuclear Fusion*, 64(2):026006, 2023. Publisher: IOP Publishing.

[25] Z.A. Xing, D. Eldon, A.O. Nelson, M.A. Roelofs, W.J. Eggert, O. Izacard, A.S. Glasser, N.C. Logan, O. Meneghini, S.P. Smith, R. Nazikian, and E. Kolemen. CAKE: Consistent Automatic Kinetic Equilibrium reconstruction. *Fusion Engineering and Design*, 163:112163, February 2021.

[26] A. Rothstein, H. Farre-Kaga, S.K. Kim, K. Erickson, and E. Kolemen. Preemptive tearing mode stabilization with multi-tasking ech in advanced tokamak plasmas. In process for publication.

[27] E. Kolemen, A.S. Welander, R.J. La Haye, N.W. Eidietis, D.A. Humphreys, J. Lohr, V. Noraky, B.G. 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. *Nuclear Fusion*, 54(7):073020, July 2014.

Appendix A: Detailed considerations on database processing and TM labelling

- I_P rampup and rampdown data is excluded, as we wished to predict flattop TMs which can be controlled by ECH steering. We wouldn't want to change ECH steering during flattop as it may affect the scenario.
- n1,2,3rms signals, the 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 labelling.
- A TM was considered to have occurred if the n1rms signal peaked above 12G for a continuous 50ms along with additional constraints on H_{98} and q_{95} to only consider H-modes 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 data was chosen to be every 20ms as this is enough time for diagnostics and EFITs to yield updated results, faster than τ_R and τ_E ensuring the profile are equilibrated, but not too fast that the model overfits to noise. Actuation such as mirror steering and NBI power adjustment will also affect the plasma on the order of $100ms$, 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 suc has q profile changes rather than actuation quantities. This would ensure the models weren't biased to predicting steered ECH suppressed TMs, but it should be a learned effect on the profiles. DIII-D also has limited variety in ECH mirror angles, making it difficult for models to learn relations between stability and mirror angle.
- The model did not have memory, which means all the diagnostic noise would appear in each prediction timestep, affecting TM predictions. We therefore apply a low pass filter to the 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.

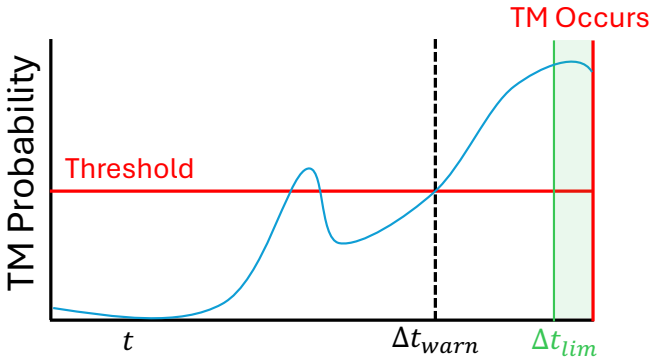

FIG. 8. Diagram of an example TM prediction. Δt_{warn} is the warning time for a TM and Δt_{lim} the time considered too late to be considered a correct prediction.

Appendix B: Event characterization details

Fig. 8 depicts an example TM prediction result to explain the event labelling 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 $\Delta t_{lim} = 100ms$ as this is 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.

Appendix C: Shapley Toy Model

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

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

With this exact formula, we can exactly see how each term, average cost $\text{Cost}_{\text{squad}}$, age Age_{avg}, and injury count $\text{Num}_{\text{injury}}$, contribute 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.

Fig. 9 shows the Shapley analysis on the win rate model using two background distributions (professional with higher costs vs amateur) and two professional teams. In a), the higher squad cost in Team 1 gives it a larger Shapley value. The higher value of squad cost ($25M$) 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 quadratic dependence on the win rate.

Team	Team 1	Team 2
Cost (millions)	$25	$24
Age (years)	22	23
Injury Count	8	5
Win rate	0.65	0.62

Context	Professional	Amateur
Cost (millions)	$15 - $30	$13 - $15
Age (years)	20 - 30	20 - 30
Injury Count	0 - 10	0 - 10

TABLE II. Left table: Example Teams 1 and 2 average cost, age, injury counts and the corresponding win rate. Right 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.

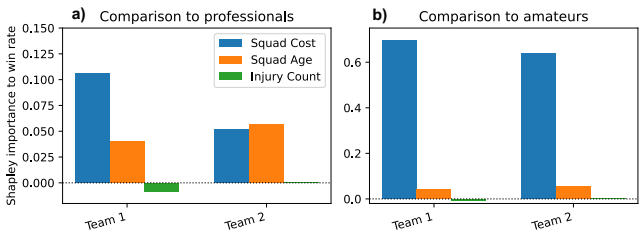

FIG. 9. Shapley values for our toy model. **a)** Shapley values between two professional teams show meaningful contributions from Cost, Age, and Injury Count, where the difference in Cost is the driving factor. **b)** Comparing to a background of Amateur teams, the cost is now the sole driving factor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

Next we see the importance of the reference distribution when comparing Fig. 9 a) to 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.