CONFERENCE PRE-PRINT

DECODING THE CAUSES OF HIGH-DENSITY DISRUPTION THROUGH INTERPRETABLE MACHINE LEARNING

[1]Chengshuo Shen, [1]Mingqiao Wen, [1]Weijie Lin, [1]Li Gao, [1]Wei Zheng, [1]Yonghua Ding and [1]Zhongyong Chen

[1]International Joint Research Laboratory of Magnetic Confinement Fusion and Plasma Physics, State Key Laboratory of Advanced Electromagnetic Technology, Huazhong University of Science and Technology, Wuhan, China

Email: shenchengshuo@hust.edu.cn

Abstract

Disruptions are catastrophic events in tokamak plasmas that can severely damage devices and compromise reliable operation, making accurate prediction and avoidance crucial for future fusion reactors. Machine learning has demonstrated strong potential in disruption prediction with high accuracy and computational efficiency, but its application is often limited by poor interpretability. These issues restrict the ability to reveal physical mechanisms and reduce the transferability of models across devices. To address this challenge, the paper focuses specifically on density-limit disruptions and develops an interpretable hierarchical classification model as a methodological attempt and validation framework. Greenwald density limit scaling provides the maximum density that plasma can reach under common conditions. However, greenwald fraction is also not the real physics of density limit disruption and cannot be used as a predictor of density limit disruption because plasma often disruption before the density limit is reached. The model is designed to predict density-limit disruptions, other disruptions, and non-disruptive discharges by excluding empirical scaling parameters and instead incorporating physics-guided features such as MARFE-related radiation asymmetries, density fluctuations, MHD activity, and plasma control system parameters. The framework employs LightGBM with a hierarchical loss function and Bayesian hyperparameter optimization to ensure both robustness and interpretability. Evaluation on 1,099 discharges from J-TEXT shows that the model achieves an overall accuracy of 96.0% for the discharges in test set. Interpretability analysis with SHAP indicates that density asymmetry and density fluctuations near 0.6a–0.7a are decisive factors in density-limit disruptions, while CIII radiation asymmetry shows a stabilizing effect. These findings confirm that the proposed method provides a feasible and interpretable approach for density-limit disruption prediction, demonstrating that physics-guided machine learning can move beyond empirical scaling to capture meaningful mechanisms and inspire more reliable disruption avoidance strategies in future tokamaks.

1. INTRODUCTION

Disruption is a catastrophic event in tokamak plasmas that requires prediction, mitigation and avoidance [1,2]. Data-driven disruption prediction has been increasingly investigated and promoted due to its outstanding performance [3–16]. However, most data-driven models are based on machine learning, which leads to a lack of interpretability. Investigating the interpretability of disruption prediction models not only validates the reliability of the models but also helps researchers understand the disruption rules that the models have learned. This helps researchers gain a deeper understanding of disruption physics, develop more suitable cross-machine models, and implement disruption avoidance strategies targeting the disruption causes.

Various interpretable or explainable disruption prediction approaches have been developed in JET [7,8], DIII-D [17], HL-2A [18], EAST [11] and J-TEXT [5,19] based on the Post-hoc interpretability methods [20]. While these approaches help validate models, their primary focus has been on verifying reliability rather than decoding the physical mechanisms underlying the predictions. Consequently, the potential of interpretability as a tool for uncovering disruption physics remains underexplored.

Understanding the causes of disruptions is not only of academic value but also of practical importance. Decoding the patterns embedded in large datasets can provide deeper insight into disruption physics, support the design of transferable cross-machine models, and enable intervention strategies targeting early precursors. However, the study of disruptions is inherently complex, making it extremely challenging to conduct interpretability research across all disruption types simultaneously. Since different types of disruptions involve distinct and possibly coupled physical mechanisms, interpretability studies focused on specific disruption categories offer a more promising route to bridge data-driven models with physical understanding.

High-density operation is particularly urgent for future tokamaks, as sustaining burning plasmas requires achieving high plasma density. However, the density limit imposes a fundamental constraint that substantially increases disruption risk. To address this, we first built a conventional model to predict all types of disruptions.

This model successfully identified the scaling relationship between plasma current and core density, analogous to the Greenwald scaling law. Nevertheless, in density-limit extension experiments with Resonant Magnetic Perturbations (RMPs), the model became overly sensitive to the line-averaged density because of its reliance on this scaling law, while neglecting important variations at the plasma boundary. This highlights the need for interpretable models capable of capturing the actual physical mechanisms driving high-density disruptions, rather than merely reproducing empirical scaling laws.

In this paper, we propose a new interpretable high-density disruption prediction model based on hierarchical classification and high-density related features. Section 2 introduces the interpretability study of high-density disruptions using the conventional model. Section 3 presents the overall model framework, including feature extraction with both non-Greenwald and physics-guided high-density features. Section 4 describes the dataset, including the data selection criteria and preprocessing procedures. Section 5 outlines the training strategy, the tri-classification approach, and evaluates the predictive performance of the proposed model. Section 6 provides interpretability analyses, validating the role of edge parameters and other physics-informed features. Finally, Section 7 summarizes the main findings and discusses their implications for disruption prediction and avoidance in future devices.

2. GREENWALD FRACTION BIAS IN DISRUPTION MODELING

An interpretable disruption predictor based on physics-guided feature extraction (IDP-PGFE) was recently developed and successfully applied on the J-TEXT tokamak [5]. This model represents a significant advancement in disruption prediction research because it combines the accuracy of modern machine learning with the interpretability provided by physics-guided features. By incorporating diagnostic signals related to magnetohydrodynamic (MHD) instabilities, radiation, density evolution, and basic plasma control system parameters, IDP-PGFE not only achieves a high degree of predictive accuracy but also offers insight into the underlying mechanisms that govern disruption onset. On J-TEXT experimental datasets, the model exhibits high predictive performance, with a true positive rate (TPR) of 97.27% and a false positive rate (FPR) as low as 5.45%. These results demonstrate that the predictor has learned sufficiently meaningful physical patterns to distinguish between disruptive and non-disruptive discharges, outperforming conventional data-driven approaches that rely solely on raw diagnostic signals.

A crucial aspect of IDP-PGFE lies in its interpretability. The model employs SHapley Additive exPlanations (SHAP)[21] to evaluate the contribution of individual physics-guided features to the prediction outcome. Detailed examination of the central line-averaged density and plasma current features revealed that the model may have effectively captured the essence of the Greenwald density limit scaling law [22]. Specifically, the SHAP analysis showed that when the plasma density approaches or exceeds a critical threshold relative to plasma current, the model assigns a higher disruptive contribution, mimicking the well-known empirical scaling that describes density limits in tokamaks. For example, when density values were above approximately 4×10^{19} m^{-3}, the contribution to disruption increased significantly, particularly when the plasma current was below 180 kA. Conversely, for plasma currents above 200 kA, the model recognized that higher density could still be tolerated without immediate disruption risk, in agreement with the higher Greenwald limit at stronger plasma currents. This finding suggests that the machine learning model, despite being trained only on experimental data, has internalized a key empirical scaling relation widely used in plasma physics.

However, while this alignment with the Greenwald law reflects the model's capacity to embed physical intuition, it also introduces potential biases that limit its interpretability and reliability in certain scenarios. The Greenwald fraction bias strongly affects the predicted results because the model tends to overemphasize density contributions in disruption assessment. A notable example is discharge #1080500, which was falsely alarmed by IDP-PGFE. In this case, the model predicted disruption primarily due to the high central density contribution. Yet, in the actual experiment, disruption was successfully avoided because the application of 3/1 and 4/1 resonant magnetic perturbations (RMPs) effectively modified the plasma radiation profile and stabilized the system against density-limit-driven termination. Interpretability analysis reveals that the model can identify variations in the contribution of the radiation profile to density-limit disruptions, but this effect is far outweighed by the dominance of the scaling relation. As a result, the predictor tends to rely on empirical density–disruption correlations rather than capturing the underlying physical stabilization mechanisms, such as impurity transport modification and turbulence suppression induced by RMPs. Thus, the contribution of the core density feature, though useful as a statistical indicator, cannot fully reflect the real physics of high-density disruption.

To address these challenges and further enhance interpretability, we propose the development of a new disruption prediction framework that explicitly removes empirical constraints such as the implicit dependence on Greenwald fraction scaling. Instead, the improved model should integrate features directly linked to high-density disruption physics. This approach would reduce over-reliance on density as a surrogate and provide a more mechanistic understanding of high-density disruptions. Such integration would significantly improve the predictor's ability to generalize across experimental scenarios and across machines.

3. THE INTERPRETABLLE HIGH-DENSITY DISRUPTION PREDICTION MODEL

We trained a hierarchical multi-label classification model based on LightGBM (LGB) to differentiate density limit disruptions, other disruption types, and non-disruptive discharges. The primary objective of this model is to identify which features can most effectively distinguish high-density disruption from all the disruption through the interpretable disruption prediction model. To mitigate potential bias induced by the Greenwald fraction, we deliberately exclude core density and plasma current as a model input feature. Instead, we preferred the physics features, such as edge transport, radiation profile (such as multifaceted asymmetric radiation from the edge, MARFE), high-density front, and MHD instabilities. The hierarchical multi-label classification model is designed to first distinguish between disruption and non-disruption events, and then further categorize disruptions into high-density disruptions and other types. To enforce hierarchical consistency, we designed a hierarchical cross-entropy loss that penalizes violations of the class structure. This framework produced a parent-class disruption prediction model and a subclass high-density disruption model.

3.1. Non-Greenwald scaling law factors

In this section, the non-Greenwald scaling law factors will be introduced. To prevent the model from simply learning empirical scaling relations, the most effective approach is to restrict such empirical scaling parameters at the input stage. Table 1 show the overview of features used in this model, the MARFE-related features are represented by ratios of diagnostic measurements obtained at different radial positions, which serve as proxies for characterizing radiation asymmetry. The numerical subscripts denote normalized minor radii (e.g., 95 corresponds to r/a = 0.95), thereby capturing radial variations in radiation behaviour. The Density Fluctuations features are introduced to reflect high-frequency perturbations that are indicative of turbulence, even though turbulence characterization itself is inherently complex. Here, the normalized gradient of line-integrated density (Den_ngrad) is evaluated specifically at r/a = 0.6 and r/a = 0.7. In addition, all fluctuation-related frequency components are filtered to exclude contributions below 20 kHz, ensuring that only high-frequency dynamics are represented. For the MHD category, Mirnov probe signals are processed to extract both frequency and amplitude information, along with the average poloidal mode number. These features are designed to capture magnetohydrodynamic activity, which is directly linked to the onset and evolution of instability precursors. Finally, the PCS-related features include toroidal field as well as plasma horizontal and vertical displacements, which are directly obtained from the plasma control system. These parameters provide essential information about the equilibrium and control of the plasma.

3.2. Improvement of Decision Tree Model Based on Hierarchical Classification

In this work, the objective is to distinguish density-limit disruptions from other types of disruptions, which naturally leads to a three-class classification problem. The three categories are density-limit disruptions (DLD), non–density-limit disruptions (NDLD), and non-disruptive discharges (ND). However, since general disruptions and density-limit disruptions are not parallel categories, a hierarchical classification approach is adopted in the machine learning framework. To provide stronger interpretability in subsequent analyses, this study does not employ deep learning models but instead adopts the decision tree–based LGB model. The main structure of the hierarchical classification model is shown in FIG. 1. FIG. 1 illustrates the overall structure of the hierarchical classification framework. The label system is organized hierarchically, where non-disruptive discharges (ND) are separated from disruptive ones in the first layer, and the second layer further distinguishes between non–density-limit disruptions (NDLD) and density-limit disruptions (DLD). The first layer (LGB Model 1) is trained with a hierarchical loss function to reduce error propagation, while its hyperparameters are optimized using Optuna with the F1-score as the evaluation target. The second layer (LGB Model 2) applies feature enhancement by incorporating the first-layer prediction as an additional input feature, which is then used alongside the original features to train the classifier. A binary log-loss is adopted as the objective function, and hyperparameter optimization is also performed with Optuna. The final prediction probabilities are derived by combining outputs from both layers, enabling consistent hierarchical decision-making. Model evaluation adopts hierarchical

accuracy and hierarchical F1-score to ensure that performance metrics respect the hierarchical label dependency and provide a comprehensive measure of classification reliability.

TABLE 1. Overview of Features Used in This Model

Physics Relation	Feature Names	Physical Meanings
MARFE	CIIIAsym (95/82/70)	Asymmetry of CIII Radiation
	HαAsym (95/82/70)	Asymmetry of Hα Radiation
	DensAsym (95/82/70)	Asymmetry of Line-Integrated Density
Density Fluctuations	Den_ngrad	Line-Integrated Density Normalized Gradient
	DenFlu_int (70,60)	Standard Deviation of Density Fluctuations
	DensFlu_fre (70,60)	Density Fluctuations Frequency
	DensFlu_amp (70,60)	Density Fluctuations amplitude
MHD	MHD_fre	Mirnov probe frequency
	MHD_amp	Mirnov probe amplitude
	MNM	Average Poloidal Mode Number
PCS	bt	Toroidal Field
	dx	Plasma Horizontal Displacement
	dy	Plasma Vertical Displacement

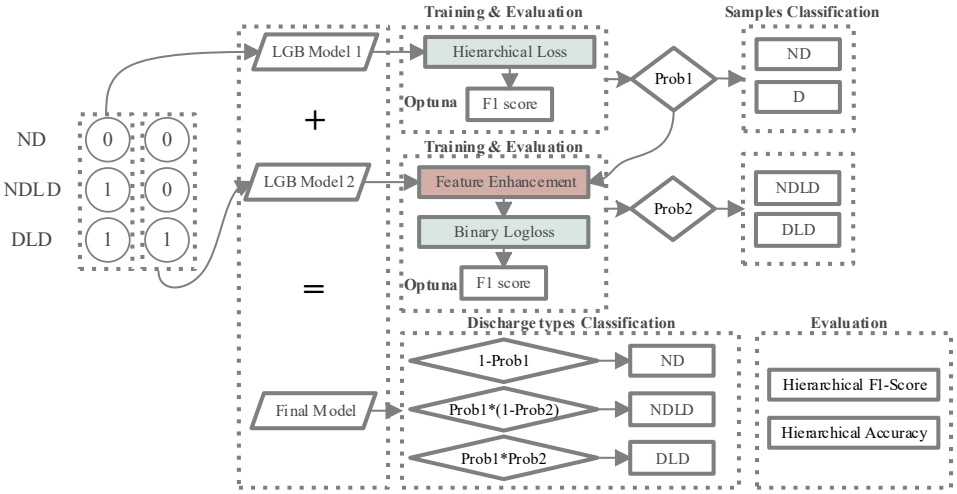

FIG. 1. Architecture of the proposed Hierarchical Decision Tree Model

For each sample, we define a label pair $(y^{(1)}, y^{(2)})$, where (0, 0) denotes a non-disruptive discharge, (1, 0) denotes a non–density-limit disruption, and (1, 1) denotes a density-limit disruption. The training objective of the first-layer model (LGB Model 1) is to predict whether a disruption will occur, corresponding to label, with the optimization target being the improved hierarchy-aware loss function (Hierarchical Loss):

$$\mathrm{L_1} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{(1)}\log(\hat{y}_i^{(1)}) + (1-y_i^{(1)})\log(1-\hat{y}_i^{(1)})\right) + \alpha\cdot\mathrm{Penalty}\cdot \tag{1}$$

The Penalty term is introduced to enhance the correctness of the higher-level classification and to prevent errors from propagating to the lower level, and is defined as:

$$\mathrm{Penalty} = \begin{cases} 1.5\times\mathrm{L_1}, & \text{if } y_i^{(1)} = 0 \text{ and } \hat{y}_i^{(1)} > 0.5 \\ \mathrm{L_1}, & \text{otherwise} \end{cases}, \tag{2}$$

where is a hyperparameter that controls the penalty strength. The second-layer model (LGB Model 2) constructs feature enhancement based on the output of the first layer, introducing the first-layer prediction as a new input feature. Together with the original features, this augmented input is used to train the second-layer classifier, which predicts the specific disruption type. During evaluation, hierarchical accuracy and hierarchical F1-score are adopted to comprehensively assess model performance, ensuring that hierarchical dependencies are respected while maximizing both classification capability and interpretability. Hierarchical accuracy requires that predictions at all levels be simultaneously correct for a sample to be considered correctly classified, which is formally expressed as:

$$\text{Hierarchical Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i^{(1)} = y_i^{(1)} \text{ and } \hat{y}_i^{(2)} = y_i^{(2)}), \tag{3}$$

where N denotes the total number of samples, $1(\cdot)$ is the indicator function that equals 1 when the prediction is entirely correct and 0 otherwise, $\hat{y}_i^{(1)}, \hat{y}_i^{(2)}$ represent the predicted labels from the first and second layers, respectively. The Hierarchical F1-Score extends the traditional F1-score, which is designed for single-level classification, to the hierarchical case. In hierarchical classification, the entire label path must be predicted correctly; therefore, the hierarchical F1-Score is defined with respect to the complete set of hierarchical label combinations. The final formula for the hierarchical F1-Score is expressed as:

$$\text{Hierarchical F1-Score} = 2 \times \frac{\text{Hierarchical Precision} \times \text{Hierarchical Recall}}{\text{Hierarchical Precision} + \text{Hierarchical Recall}}. \tag{4}$$

4. TRAINING AND PERFORMANCE

4.1. Database

In this paper, the database search covered a total of 38,000 discharge experiments conducted on the J-TEXT device between January 16, 2017, and December 30, 2022. The selection criteria are that the diagnostics of all the discharges should be available for extracting feature in table 1 and the discharge should at least maintain 0.2 second. From these, 1,099 discharges that met the selection criteria were chosen. As shown in Table 2, these discharges were randomly divided into training, validation, and test sets in a 7:1:2 ratio. A density-limit disruption was defined as a high-density discharge in which the plasma density reached at the time of disruption. In the identification process for both types of disruptive discharges, non-spontaneous disruptions caused by SMBI (Supersonic Molecular Beam Injection) and SPI (Shattered Pellet Injection) were excluded.

TABLE 2. Split of datasets

	Shot No. of ND	Shot No. of NDLD	Shot No. of DLD
Training	262	254	253
Validation	38	36	36
Test	75	73	72

4.2. Performance

The One-vs-Rest ROC is a multi-class evaluation method whose core idea is to treat each class in turn as the positive class while merging the remaining classes as the negative class, thereby constructing multiple binary ROC curves. Each binary ROC curve is generated by calculating the FPR and TPR under different threshold values. As shown in Fig. 2, the One-vs-Rest ROC curves demonstrate the classification performance of the hierarchical model across the three categories. The ROC curve for ND reaches an AUC of 0.91, while NDLD and DLD achieve higher AUC values of 0.95 and 0.96, respectively. In addition, the macro-average ROC curve yields an overall area of 0.94, indicating that the model maintains strong and balanced predictive capability for all three classes.

In the disruption prediction task, the prediction is transformed from classifying individual samples to generating time-evolving predictions for an entire discharge. In traditional binary classification, the disruptive probability output can be distinguished using a 0–1 threshold. However, for a three-class model, threshold determination must simultaneously account for all three categories. To address this, we selected a threshold selection criterion based on the G-Mean metric:

$$\text{G-Mean} = \sqrt{\text{TPR} \times (1 - \text{FPR})}. \tag{5}$$

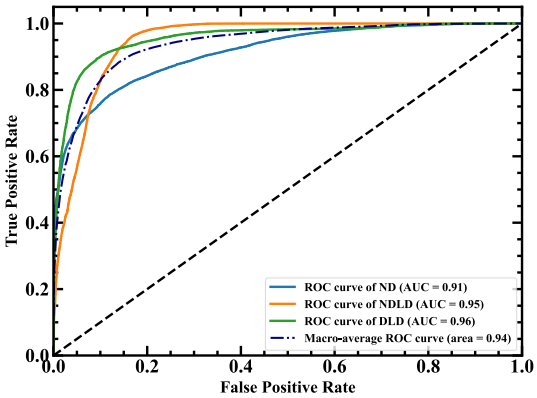

FIG. 2. One-vs-Rest ROC Curves of the hierarchical classification model for the three sample categories

The confusion matrix generated from 224 discharges in the test set is shown in Fig. 3. From the confusion matrix, it can be observed that the model achieves consistently high classification performance across all three discharge types. Specifically, the traditional disruption prediction accuracy (disruptive vs. non-disruptive) reaches 96.0% (215/224). The recognition accuracy for non-disruptive discharges (ND) is 96.1% (73/76), with almost no false alarms. For non–density-limit disruptions (NDLD), the recognition accuracy is 91.9% (68/74), with a small number of false positives and false negatives. For density-limit disruptions (DLD), the recognition accuracy is 87.8% (65/74), where most errors arise from confusion with other types of disruptions.

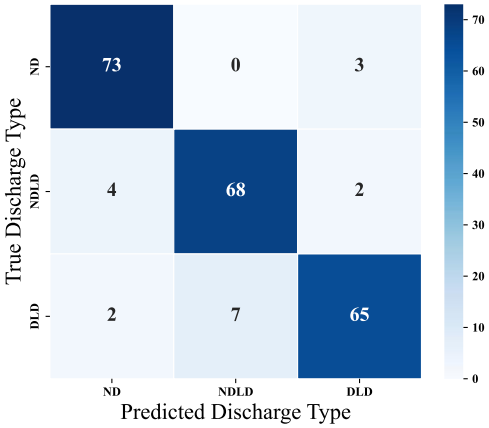

FIG. 3. Confusion matrix of prediction results for the three discharge categories

5. INTERPRETABILITY STUDY OF DENSITY LIMIT DISRUPTION PREDICTION

The hierarchical classification model is capable of successfully predicting disruptions and distinguishing density-limit disruptions even without the Greenwald fraction as an input. This demonstrates that the model does not rely on empirical scaling relations but can instead learn the underlying patterns from features derived through existing physical understanding of disruptions. This section employs SHAP-based interpretability analysis to uncover the rules identified by the model and to provide insights that may inspire future physics studies.

FIG. 4 presents the SHAP beeswarm plots of MARFE-related features, where indices 1, 2, and 3 correspond to diagnostic positions at 0.95a, 0.82a, and 0.7a on the high- and low-field sides, respectively. Panel (a) illustrates how the model distinguishes disruptive from non-disruptive discharges, while panel (b) shows how density-limit disruptions are separated from other disruption types. The results indicate that signals at 0.82a, and 0.7a carry greater importance than those at 0.95a, suggesting that MARFE formation at the very edge may not immediately exert a decisive influence on disruption onset. Furthermore, stronger density asymmetry is consistently associated with a higher probability of being classified as a density-limit disruption. By contrast, the contributions of CIII radiation asymmetry and density asymmetry exhibit opposite trends: while density asymmetry strengthens the disruption prediction, CIII asymmetry tends to mitigate it, implying a competing role between edge density gradients and radiation asymmetries in the disruption process.

FIG. 5 shows the beeswarm plots of SHAP values for density fluctuation-related (turbulence-related) features, where indices 1 and 2 correspond to positions at 0.7a and 0.6a, typically near the location where density fluctuation activity is enhanced around the q = 2 surface. In Layer 1, the results indicate that stronger density fluctuations and

steeper density gradients significantly increase the probability of disruption, reflecting the destabilizing influence of turbulence-driven transport. In Layer 2, which differentiates density-limit disruptions from other disruption types, the model decisions are primarily based on whether density fluctuations intensify and whether the density gradient rises further. Moreover, density fluctuation-related features from the 0.6a show stronger discriminative capability, suggesting that inward-shifted density fluctuation plays a more decisive role in driving the plasma toward density-limit disruptions.

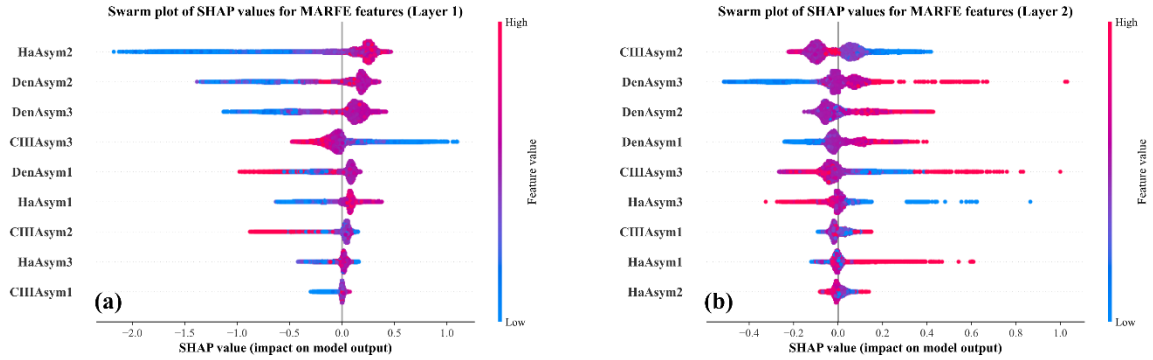

FIG. 4. Beeswarm Plot of Global SHAP Contributions for MARFE-Related Features

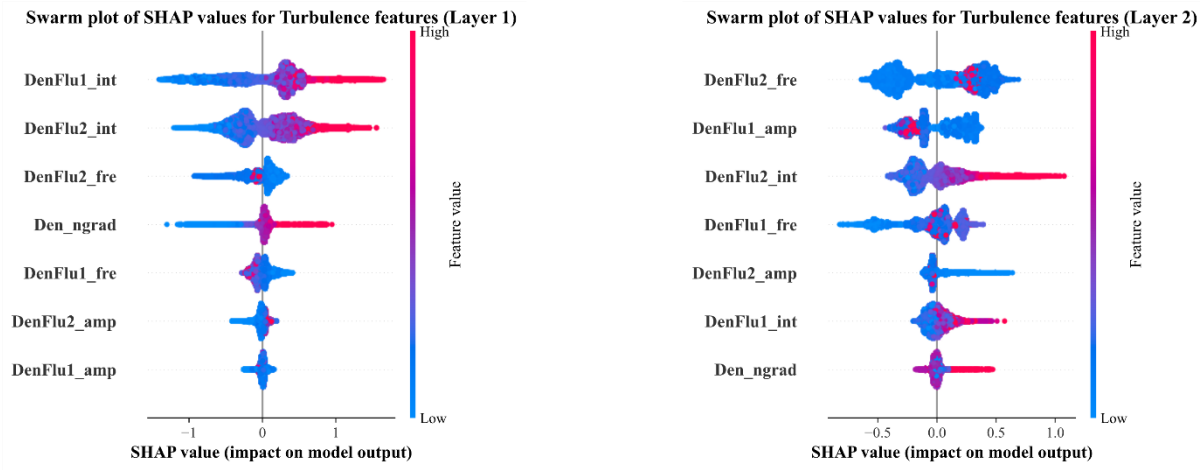

FIG. 5. Beeswarm Plot of Global SHAP Contributions for density fluctuation-related Features

6. SUMMARY

In this work, we developed an interpretable hierarchical disruption prediction model to separate DLD from other NDLD and ND. Unlike conventional models that implicitly depend on the Greenwald fraction, our approach deliberately excludes such empirical scaling parameters and instead incorporates physics-guided features, including MARFE-related asymmetries, density fluctuation measures, MHD activity, and PCS signals. The model was implemented using LightGBM with a hierarchical classification framework, a custom hierarchical loss, and Optuna-based Bayesian hyperparameter optimization, ensuring both robustness and interpretability.

The proposed model demonstrates strong predictive capability, achieving a macro-average AUC of 0.94 of samples and an overall disruption prediction accuracy of 96.0% on J-TEXT experimental data. Interpretability analysis using SHAP shows that edge density asymmetries and turbulence near 0.6a to 0.7a are decisive factors in density-limit disruptions, while CIII radiation asymmetry tends to have a stabilizing effect. These findings demonstrate that physics-guided machine learning can move beyond empirical scaling laws, providing both reliable disruption prediction and valuable physical insight that can inform avoidance strategies in future tokamaks.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the help from J-TEXT team. This work was supported by National Key R&D Program of China under Grant (No. 2024YFE03230100 and No. 2022YFE03040004), Hubei International Science and Technology Cooperation Projects (No. 2022EHB003), Natural Science Foundation of Wuhan (No.

2024040701010040) and by National Natural Science Foundation of China (NSFC) under Project Numbers Grant (No. 12375219 and No. T2422009).

REFERENCES

[1] Boozer, A. H. Theory of tokamak disruptions. *Phys. Plasmas* **19**, 058101 (2012).
[2] Lehnen, M. *et al.* Disruptions in ITER and strategies for their control and mitigation. *J. Nucl. Mater.* **463**, 39–48 (2015).
[3] Kates-Harbeck, J., Svyatkovskiy, A. & Tang, W. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature* **568**, 526–531 (2019).
[4] Vega, J., Murari, A., Dormido-Canto, S., Rattá, G. A. & Gelfusa, M. Disruption prediction with artificial intelligence techniques in tokamak plasmas. *Nat. Phys.* **18**, 741–750 (2022).
[5] Shen, C. *et al.* IDP-PGFE: an interpretable disruption predictor based on physics-guided feature extraction. *Nucl. Fusion* **63**, 046024 (2023).
[6] Shen, C. *et al.* Cross-tokamak disruption prediction based on domain adaptation. *Nucl. Fusion* **64**, 066036 (2024).
[7] Aymerich, E. *et al.* A self-organised partition of the high dimensional plasma parameter space for plasma disruption prediction. *Nucl. Fusion* **64**, 106063 (2024).
[8] Pau, A. *et al.* A machine learning approach based on generative topographic mapping for disruption prevention and avoidance at JET. *Nucl. Fusion* **59**, 106017 (2019).
[9] Zhu, J. X. *et al.* Hybrid deep-learning architecture for general disruption prediction across multiple tokamaks. *Nucl. Fusion* **61**, 026007 (2020).
[10] Guo, B. H. *et al.* Disruption prediction on EAST with different wall conditions based on a multi-scale deep hybrid neural network. *Nucl. Fusion* **63**, 094001 (2023).
[11] Hu, W. H. *et al.* Real-time prediction of high-density EAST disruptions using random forest. *Nucl. Fusion* **61**, 066034 (2021).
[12] Yang, Z. *et al.* A disruption predictor based on a 1.5-dimensional convolutional neural network in HL-2A. *Nucl. Fusion* **60**, 016017 (2019).
[13] Zheng, W. *et al.* Hybrid neural network for density limit disruption prediction and avoidance on J-TEXT tokamak. *Nucl. Fusion* **58**, 056016 (2018).
[14] Zheng, W. *et al.* Disruption prediction for future tokamaks using parameter-based transfer learning. *Commun. Phys.* **6**, 181 (2023).
[15] Rea, C., Montes, K. J., Erickson, K. G., Granetz, R. S. & Tinguely, R. A. A real-time machine learning-based disruption predictor in DIII-D. *Nucl Fusion* (2019).
[16] Murari, A. *et al.* A control oriented strategy of disruption prediction to avoid the configuration collapse of tokamak reactors. *Nat. Commun.* **15**, 2424 (2024).
[17] Rea, C., Montes, K. J., Pau, A., Granetz, R. S. & Sauter, O. Progress Toward Interpretable Machine Learning–Based Disruption Predictors Across Tokamaks. *Fusion Sci. Technol.* (2020).
[18] Yang, Z. *et al.* In-depth research on the interpretable disruption predictor in HL-2A. *Nucl. Fusion* **61**, 126042 (2021).
[19] Ding, Y. *et al.* Overview of the recent experimental research on the J-TEXT tokamak. *Nucl. Fusion* **64**, 112005 (2024).
[20] Madsen, A., Reddy, S. & Chandar, S. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Comput. Surv.* **55**, 1–42 (2023).
[21] Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
[22] Shen, C. *et al.* In-depth research on the interpretability and fewer data learning for the disruption predictor in J-TEXT. 49th EPS Conference on Contr. Fusion and Plasma Phys, P4.016[R], Bordeaux: EPS (2023).