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

[1]Chengshuo Shen, [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 Engineering and Technology, School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan, China

Email: shenchengshuo@hust.edu.cn, zhengwei@hust.edu.cn

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]. However, most data-driven models are based on machine learning, which leads to a lack of interpretability. Post-hoc interpretability methods [4] can effectively demystify black-box models. To date, most interpretable disruption prediction approaches show the intrinsic relationships between the plasma state and disruption, or the contributions of features that lead to disruption [5,6]. However, their primary objective is to validate the reliability of the models, rather than to decode the underlying patterns mined from extensive datasets. Decoding the causes of disruptions will helps researchers gain a deeper understanding of disruption physics, develop more suitable cross-machine models, and intervene the disruption precursors. High-density operation is crucial for ITER and DEMO, but the density limit imposes a fundamental constraint, increasing disruption risks. In this paper, we focus on the interpretability study of the high-density disruption. Initially, we built a conventional model to predict all types of disruptions. It successfully identified the scaling relationship between plasma current and core density, which analogous to Greenwald scaling law. However, experiments on the J-TEXT found that the scaling relationship could lead to false alarms. Consequently, we developed a new model that can capture the underlying physical mechanisms of high-density disruptions rather than merely the scaling law.

2. GREENWALD FRACTION BIAS IN DISRUPTION MODELING

An interpretable disruption predictor based on physics-guided feature extraction (IDP-PGFE) was recently developed on the J-TEXT [7]. The model exhibits high predictive performance with high true positive rate (TPR = 97.27%) and low false positive rate (FPR = 5.45%). Detailed examination of the roles of central chord averaged density and plasma current by SHAP indicates that the model may have effectively captured the Greenwald scaling law [8]. However, the Greenwald fraction bias will strongly affect the predicted result. The discharge # 1080500 was false alarmed because of the high contribution of n_e. Actually, the application of 3/1 and 4/1 Resonant Magnetic Perturbation (RMP) reduced the contribution of the radiation profile avoid the high-density disruption. Therefore, the contribution of the core density could not reflect the real physics of high-density disruption, just the empirical relationship. To address this, we propose a new disruption prediction model, which removes empirical constraints, integrating impurity radiation and turbulence transport related features for enhanced the interpretability of high-density disruption.

3. HIERARCHICAL MULTI-LABEL DISRUPTION PREDICTION MODEL FOR HIGH-DENSITY DISRUPTION

We trained a hierarchical multi-label classification model based on LightGBM 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. If the core density is higher than $0.6n_G$ at the disruption time, the discharge will be treated as the high-density disruption. To mitigate potential bias induced by the Greenwald fraction, we deliberately exclude core density as a model input feature. Instead, we preferred the physics features, such as edge transport [9,10], radiation profile (such as multifaceted asymmetric radiation from the edge, MARFE), high-density front [11], 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.

4. INTERPRETABILITY ANALYSIS WITHOUT GREENWALD FRACTION BIAS

The strong performance of the hierarchical multi-label classification model indicates that it has successfully identified the underlying patterns distinguishing high-density disruption from the other disruption types without Greenwald fraction bias. Notably, this model correctly identified shot #1080550 as a non-disruptive discharge. This demonstrates that the Greenwald fraction, as an empirical scaling law, can mislead disruption prediction models. Therefore, design the disruption models that inherently avoid such biases is necessary. The interpretability analysis shows that the parent-class disruption prediction model mainly through MHD instabilities and radiation profile related features. In the subclass high-density disruption model, vertical displacement emerged as the most critical feature, with lower values correlating with an increased risk of high-density disruption. On J-TEXT, the vertical displacement reflects the position of the current centroid, as measured by magnetic diagnostics. The contribution of vertical displacement may reflect significant differences in the current density profiles between high-density disruption and other disruption types. The impact of vertical displacement in high-density disruption and other disruption types has been overlooked in previous experiments.

5. CONCLUSION

In this study, we have developed an interpretable framework for disruption prediction to decode causes of high-density disruption. Our initial investigations using a conventional model found that while the Greenwald scaling law effectively captures the empirical relationship between plasma current and core density, its overreliance can lead to false alarms. To mitigate Greenwald fraction bias, we developed a hierarchical multi-label classification model that excludes the core density input. The interpretability analysis suggests that differences in the current density profile, as indicated by vertical displacement measurements, which has been overlooked in previous studies.

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. Physics of Plasmas 19, 058101 (2012).
[2]. Lehnen, M. et al. Disruptions in ITER and strategies for their control and mitigation. Journal of Nuclear Materials 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]. Madsen, A., Reddy, S. & Chandar, S. Post-hoc Interpretability for Neural NLP: A Survey. ACM Comput. Surv. 55, 1–42 (2023).
[5]. Rea, C., Montes, K. J., Pau, A., Granetz, R. S. & Sauter, O. Progress Toward Interpretable Machine Learning–Based Disruption Predictors Across Tokamaks. Fusion Science and Technology (2020).
[6]. 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).
[7]. Shen, C. et al. IDP-PGFE: an interpretable disruption predictor based on physics-guided feature extraction. Nucl. Fusion 63, 046024 (2023).
[8]. 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, Bordeaux (2023).
[9]. Ke, R. et al. Electrode biasing maintains the edge shear layer at high density in the J-TEXT tokamak. Nucl. Fusion 62, 076014 (2022).
[10]. Long, T. et al. The role of shear flow collapse and enhanced turbulence spreading in edge cooling approaching the density limit. Nucl. Fusion 64, 066011 (2024).
[11]. Shi, P. et al. Observation of the high-density front at the high-field-side in the J-TEXT tokamak. Plasma Phys. Control. Fusion 63, 125010 (2021).