In-depth research on the interpretability and fewer data learning for the disruption predictor in J-TEXT

C. Shen[1], W. Zheng[1], B. Guo[2, 3], D. Chen[2], X. Ai[1], F. Xue[1], Y. Zhong[1], N. Wang[1], B. Shen[2], B. Xiao[2], Y. Ding[1], Z. Chen[1], Y. Pan[1] and J-TEXT team[†]

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

[2] *Institute of Plasma Physics, Chinese Academy of Sciences, Hefei, China*

[3] *College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, China*

Email address of the main author: shenchengshuo@hust.edu.cn, zhengwei@hust.edu.cn

[†] SEE THE AUTHOR LIST OF "N. WANG *ET AL* 2022 ADVANCES IN PHYSICS AND APPLICATIONS OF 3D MAGNETIC PERTURBATIONS ON THE J-TEXT TOKAMAK, *NUCL. FUSION* **62** 042016"

Abstract

The future tokamak disruption predictors will need to be reliable and have the ability to learn from fewer data. Utilizing existing knowledge of disruption physics and tokamak discharge could both increase the interpretability and reduce the data demand for a disruption predictor. The Interpretable Disruption Predictor based on Physics-Guided Feature Extraction (IDP-PGFE) has been successfully applied on J-TEXT and already show its interpretability and low data learning ability. The interpretability analysis shows that IDP-PGFE learned certain physical laws related to disruptions. We in-depth investigated the contributions of the variations between features to disruptions and discovered that IDP-PGFE may have learned the Greenwald density limit scaling law. Meanwhile, IDP-PGFE can achieve enough good performance when the number of disruptive discharges is as low as around 20 discharges in J-TEXT. However, as the number of disruptive discharges drops to around 10 discharges, only the data from one tokamak is unable to train an acceptable model. Therefore, we adapted the domain adaption algorithm CORrelation Alignment (CORAL) for fewer data learning. Benefitting from PGFE and CORAL, a cross-machine disruption prediction with acceptable performance from J-TEXT to EAST could be achieved by using fewer discharges from EAST and thousands of discharges from J-TEXT. In general, physics-guided approach could provide the disruption predictor better interpretability, thus making it more reliable. The preliminary results of cross-machine disruption prediction also demonstrate the ability of physics-guided approaches to train models using fewer data from the target tokamak.

1. Introduction

Disruptions in future tokamaks need to be avoided, prevented and mitigated from the threats effects like thermal loading, electromagnetic loading and runaway electrons[1]. In recent years, disruption prediction, especially data-driven disruption prediction, has provided a

solution that can be used to trigger disruption avoidance, prevention and mitigation strategies[2–6]. However, most disruption predictors rely on a large amount of data from tokamak devices to train usable disruption prediction models. The cost of achieve such experimental data from future tokamaks is very high, therefore, the future tokamak disruption predictors will need to have the ability to learn from fewer data. Meanwhile, interpretable analysis can help researchers better understand the knowledge learned by the disruption predictor, ensuring the reliability of the model, and providing clearer directions for its extension and improvement[7,8].

To reduce the demand for training data and increase interpretability in machine learning models, it is necessary to improve the learning efficiency of the models. Introducing domain knowledge inductive bias into the model can help machine learning models learn domain-specific knowledge easier. IDP-PGFE[9] in J-TEXT has achieved true positive rate (TPR) ~ 90% and false positive rate (FPR) ~ 10% by utilizing only 20 disruption discharges and hundreds of non-disruption discharges. The interpretability analysis of this predictor demonstrates that the model has learned the knowledge of disruptions on the J-TEXT tokamak. Furthermore, the low-data learning capability of the model is attributed to the feature extractor called Physics-Guided Feature Extraction (PGFE). However, as the number of disruptive discharges drops to around 10 discharges, only the data from one tokamak is unable to train an acceptable model. Transferring a disruption predictor from an existing tokamak to a new tokamak is also an effective way to reduce the data required for the new tokamak[2,4]. This is a very typical problem addressed by transfer learning[10]. A theorem of transfer learning suggests in order to achieve better performance in the target domain, it requires better performance in the source domain and smaller differences between the source and target domains[11]. IDP-PGFE can be regarded as a source domain model with relatively good performance. Methods in domain adaptation[12] can effectively reduce the discrepancy between the source domain and the target domain.

In this research, we present that IDP-PGFE could even learn the Greenwald density limit scaling law in section 2. Then the preliminary result of using CORAL[13] to predict disruption with fewer data on EAST will be shown in section 3. Section 4 is a brief summary.

2. The in-depth interpretability analysis of IDP-PGFE

The outstanding performance of IDP-PGFE and its initial interpretability analysis has been made in previous research[9]. IDP-PGFE with physics-guided features has a TPR of 97.27% and FPR of 5.45%. The interpretability study using SHapley Additive exPlanations (SHAP)[14] provides an understanding of J-TEXT disruption and matches the prior comprehension of disruption and the disruption process.