CONFERENCE PRE-PRINT

AUGMENTING THE EXTRAPOLATION CAPABILITY OF DISRUPTION PREDICTION TO EXTENDED PARAMETER REGIMES BY PREDICT-FIRST NEURAL NETWORK

Z. Y. Yang[1], *W. L. Zhong[1], J. Y. Li[1], Y. H. Chen[1], D. Li[1], J. Z. Zhang[1], X. Fan[1], B. Li[1], Y. P. Zhang[1], Y. B. Dong[1], J. Artola[2], T. F. Sun[1], Z. H. Xu[1], R. S. Qiu[1], X. Sun[1] and J. R. Wen[1]
1. Southwestern Institute of Physics, Chengdu 610041, China
2. ITER Organization, route de Vinon sur Verdon, 13067 St Paul Lez Durance, France
Email: zhongwl@swip.ac.cn

Abstract

The prediction, mitigation, and avoidance of disruptions are critical prerequisites for ensuring the safe operation of future fusion reactors. While machine learning-based disruption prediction techniques have achieved high accuracy in recent years, their application to a new device demands strong extrapolation capability across parameter regimes. This capability is essential to maintain robust algorithm performance as the operational space of the device expands, while it is also a main shortcoming for most machine learning techniques. This study evaluates the extrapolation capability of disruption prediction algorithm using experimental data from early campaigns on the HL-3 tokamak, yielding three key findings. Firstly, the extrapolation performance of the standard deep learning algorithm is not good. However, the Predict-First Neural Network (PFNN) significantly enhances the performance. Secondly, the algorithm's accuracy exhibits distinct variations when extrapolating based on electromagnetic parameters (plasma current I_p, toroidal magnetic field B_t, safety factor q_{95}) compared to energy-related parameters (stored energy W_e, normalized beta β_N). Finally, after rigorously accounting for the impact of parameter regime extrapolation and adjusting the algorithm deployment strategy, deep learning proved effective in protecting the device during HL-3's high-current and high-beta commissioning phases. This study provides valuable experience for implementing disruption prediction algorithms on new devices and is expected to offer practical reference for the initial operation of future fusion facilities like ITER.

1. INTRODUCTION

Disruption, an abrupt termination of plasma discharge triggered by the development of instabilities or loss of control, poses a significant threat to future fusion reactors. The resulting consequences, including thermal loads, electromagnetic forces, and runaway electron beams, are predicted to reach unacceptable levels in reactor-scale devices[1]. Therefore, reliable prediction, coupled with effective mitigation and avoidance strategies, is imperative to minimize disruption damage[2, 3].

Machine learning, particularly deep learning, has demonstrated considerable promise in disruption prediction over recent years[4-9]. These techniques have achieved not only high accuracy but have also been reinforced by complementary tools tailored for the application in future devices, such as transfer learning, interpretability analysis, and anomaly detection[10-12]. Significant efforts are underway to integrate these methodologies into a comprehensive, systematic solution for disruption management in future fusion power plants.

The HL-3 tokamak, the largest tokamak currently operated in China, is designed to operate within a plasma current (I_p) range of 2.5-3 MA. Since its initial plasma in 2020[13], HL-3 has progressively pushed towards higher performance parameters, recently achieving H-mode operation at 1.5 MA plasma current[14]. As it steadily approaches its design goals, HL-3 inevitably encounters the issue of disruption risks. This makes it an ideal "test bed" for disruption prediction algorithms, providing invaluable experience for future reactor operation.

The development of disruption prediction algorithms for HL-3 is described in our previous paper[15], where a critical issue is noticed. Previous cross-machine disruption prediction studies often treated a "future device" as an environment with scarce training data but a relatively stable underlying data distribution. This differs subtly from the reality of a newly constructed device like HL-3. Such a device accumulates substantial data during its initial low-parameter commissioning phase; however, its operational parameter space continuously expands. This dynamic expansion presents a unique challenge for disruption prediction algorithms, demanding strong extrapolation capabilities. While approaches like scenario adaptive learning have been proposed to address aspects

of this challenge[16, 17], there remains a notable lack of direct, quantitative performance evaluation under conditions of parameter space expansion. This study aims to address this crucial gap.

The remainder of this paper is structured as follows. Section 2 describes the experimental dataset used and details its partitioning into training and testing sets, guided by HL-3's high-parameter commissioning strategy. Section 3 outlines the algorithm architectures and compares the extrapolation capabilities of baseline model and the Predict-First Neural Network (PFNN). It specifically analyzes performance differences when extrapolating based on electromagnetic configuration-related parameters (I_p, B_t, q_{95}) versus energy-related parameters (W_e, β_N). Building on these insights, Section 4 details the deployment and application of the refined algorithms during HL-3's high-current and high β_N commissioning, presenting demonstrative results. Finally, Section 5 summarizes the key findings and conclusions.

2. DATASET DESCRIPTION

2.1 HL-3 operational space

HL-3 has conducted five experimental campaigns to date. After excluding discharges with a maximum plasma current below 100 kA, a total of 3819 shots were available for this study. Throughout these campaigns, HL-3's primary commissioning objectives focused on enhancing plasma current, normalized beta, ion temperature, and the fusion triple product. This study mainly concentrates on investigating the extrapolation of algorithms across plasma current and normalized beta.

Figure 1 depicts the distribution of maximum I_p versus average B_t across the entire dataset, alongside the distribution of maximum I_p versus maximum β_N. The figure reveals that a substantial number of shots were dedicated to progressively ramping I_p towards 1.6 MA. Additionally, significant experimental effort was directed at achieving high β_N discharges at plasma current around 300 kA, 500 kA, and 700 kA.

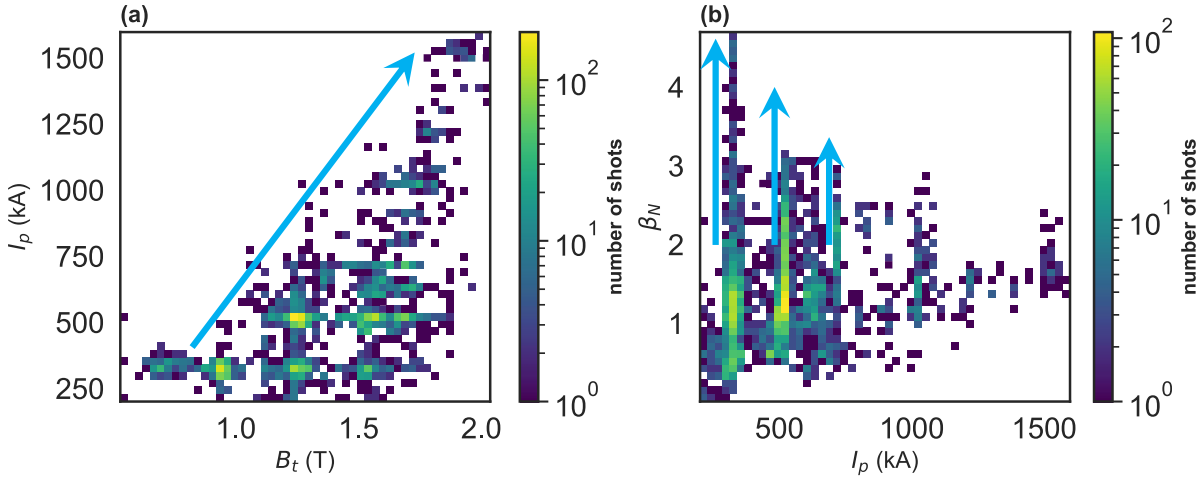

Figure 1 The diagram of HL-3's operational parameter space, which illustrates two of the device's primary commissioning objectives, as indicate by blue arrow. Firstly, enhancing the plasma current (I_p) up to 1.6MA. Secondly, enhancing normalized beta (β_N) at specific plasma current levels, namely 300 kA, 500 kA, and 700 kA.

2.2 Dataset organization

Reflecting HL-3's two primary commissioning paths, high I_p and high β_N operation, the dataset is partitioned based on maximum I_p, average B_t, minimum q_{95}, maximum W_e, and maximum β_N by shot. This partitioning enables the examination of algorithm performance when extrapolating along each individual parameter dimension.

Figure 2 presents the statistical distributions of these five parameters across the full dataset. For each parameter, the dataset was divided into four segments. The largest segment was allocated for training and validation, while the remaining three segments served as distinct test sets. Crucially, the parameter ranges covered by the test sets

are completely different from those in the validation set. This design isolates the impact of operational regime shifts on algorithm performance.

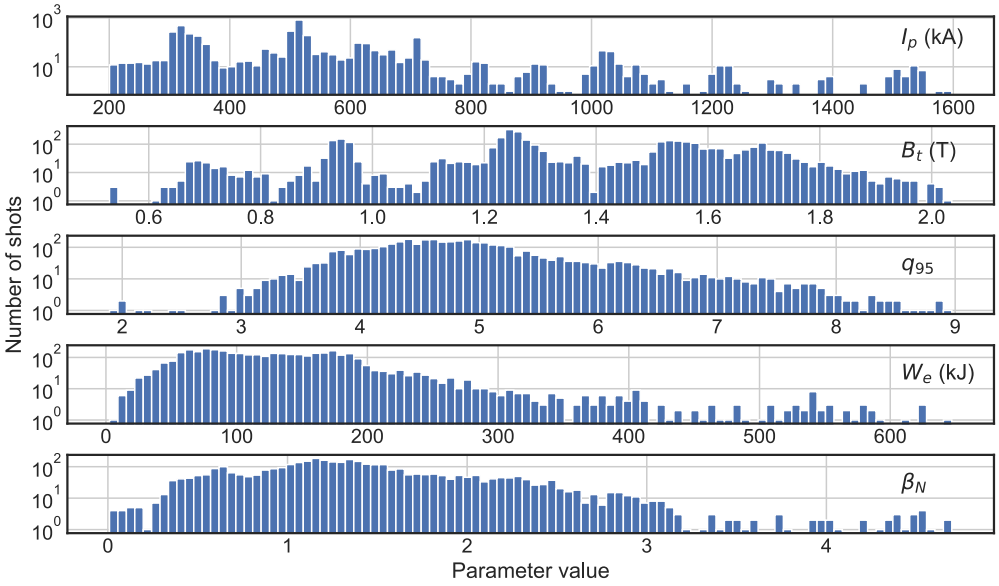

Fig 2 The distribution histograms of key plasma parameters across the dataset. The horizontal axis represents the value of each parameter, while the vertical axis indicates the number of experimental shots within each bin. Logarithmic scaling is applied to the vertical axis to effectively visualize sparsely populated regions within the parameter space.

For nearly every parameter, the shots within the parameter ranges targeted for extrapolation are significantly sparser. To mitigate the confounding effect of sample size on model performance, the partitioning thresholds for all five parameters were carefully set to achieve an similar shot distribution across the four segments. The exact shot counts and parameter thresholds are detailed in Table 1.

Beyond the partitioning scheme designed primarily for research purposes, the disruption prediction algorithm was also deployed operationally during HL-3's 1.6 MA I_p commissioning and high β_N commissioning for risk mitigation. In these practical scenarios, adhering to the principle of maximizing the accuracy of the deployed version, all available data were utilized for training and validation. The corresponding dataset information is given in Section 4.

Table 1 The parameter scope and number of shots within each subdatasets.

Parameter	Range	Number of shots	Dataset ID
I_p (kA)	0~520	2570	1
	520~700	748	2
	700~1000	281	3
	1000~1700	220	4
B_t (T)	0~1.55	2652	1
	1.55~1.7	769	2
	1.7~1.78	253	3
	1.78~2.2	141	4
q_{95}	4.5~10	2495	1
	4~4.5	870	2
	3.7~4	292	3
	1.8~3.7	159	4
W_e (kJ)	0~160	2545	1
	160~220	871	2
	220~300	252	3
	300~700	148	4

β_N	0~1.5	2485	1
	1.5~2.1	800	2
	2.1~2.5	324	3
	2.5~5	207	4

2.3 Label of disruption

This study adopts the data labeling strategy established in our prior research:

— Data within 30 ms prior to disruption are labelled as "disruptive".

— Data between 30 ms and 200 ms prior to disruption are assigned a "fuzzy" label. This accounts for the inherent variability in the timing of disruption precursors across different shots. A hard transition could introduce wrong labels and negatively impact model training.

— Data more than 200 ms prior to disruption and data from non-disruptive shots are labelled as " non-disruptive".

It is important to note that the HL-3 dataset contains a significant number of shots where the Disruption Mitigation System (DMS) are triggered. Labelling these shots presents a significant challenge. Since it is hard to definitively determine whether the plasma was in a normal state or already exhibiting disruption precursors before the DMS injection. To address this issue, this study truncates the final 200 ms of data from these shots. The remaining data from these shots is then treated as non-disruptive shots. This approach partially mitigates the labelling ambiguity problem. Effectively utilizing data involving DMS remains an important open question for future research, particularly in the context of future fusion reactors.

3. ALGORITHMS AND EXTRAPOLATION TESTING RESULTS

3.1 Predict-first neural network

Regarding the input signal selection and neural-network structure, this study follows the methodology established in our previous work[15]. The implementation details are omitted here for brevity. This section focuses specifically on the Predict-first Neural Network (PFNN) architecture central to our investigation.

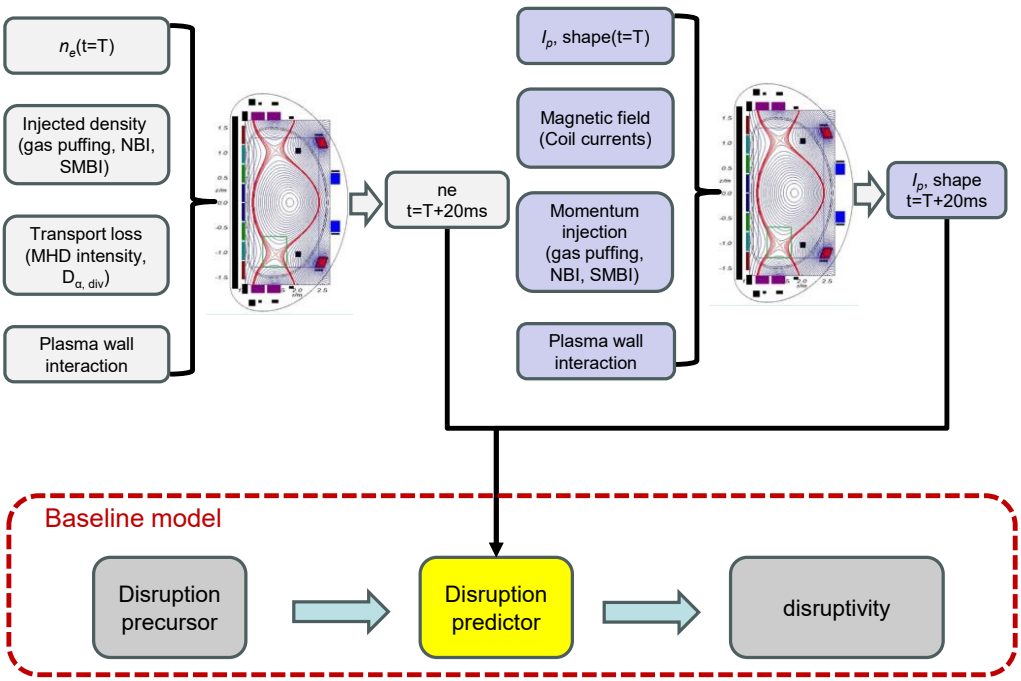

Fig 3 the overall architecture of the PFNN model.

Figure 3 illustrates the overall architecture of the PFNN model. The core concept of PFNN involves a two-stage process. Firstly, a predictive model is trained to forecast the temporal evolution of plasma current, magnetic configuration, and plasma density in 20ms. Secondly, the difference between the simulated evolution and the actual experimental result is then computed. This simulation-experiment discrepancy serves as the primary input feature for the subsequent disruption predictor. It is important to note that when extrapolating to new operational regimes, the plasma evolution predictor itself may also suffer from performance degradation due to the unfamiliar parameter space. However, this limitation can be effectively addressed by fine-tuning the predictor using experimental data from 1~2 shots within the new target regime, as shown in [18].

3.2 Analysis on extrapolation testing results

Figure 4 presents the performance of both the baseline model and the PFNN algorithm across various test datasets. The performance is evaluated by area under receiver-operator characteristic curve (ROC curve), namely AUC, as in most related research. Two principal conclusions can be inferenced.

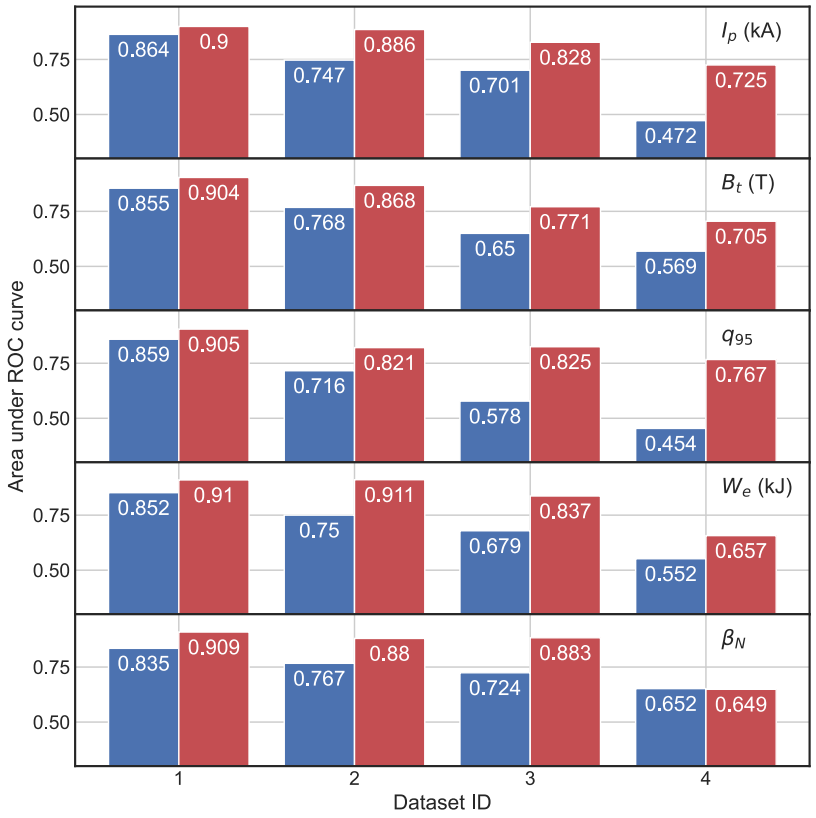

Fig 4 the AUC of both the baseline model(blue) and the PFNN algorithm(red) across various datasets.

Enhanced extrapolation capability of PFNN: The PFNN architecture significantly enhances the extrapolation capability of the disruption prediction algorithm. While the baseline algorithm exhibits a rapid decline in accuracy when extrapolating to new regimes, the PFNN maintains substantially stable performance over a wider range or demonstrates a markedly slower degradation rate. This phenomenon arises because the baseline model predominantly captures correlations between numerical values of plasma parameters and disruption risk. Such correlations are often highly regime-specific and prone to failure under significant parameter shifts. In contrast, PFNN shifts the focus towards monitoring the discrepancy between predicted plasma evolution and actual experimental measurements. Provided the evolution predictor remains reliable, this discrepancy-based approach offers a more universally applicable indicator of underlying instability across diverse operational regimes. On the other hand, the effectiveness of PFNN's approach inherently depends on the extrapolation capability of the plasma evolution predictor itself. Our related research suggests that fine-tuning this evolution predictor for a new regime appears significantly more tractable than adapting the disruption predictor directly. Furthermore, this evolution predictor could also be implemented using physical simulation codes.

Relationship between algorithm performance and extrapolated parameter: The degradation pattern of algorithm performance exhibits a marked distinction when extrapolating based on electromagnetic parameters versus energy-related parameters. For electromagnetic parameters (I_p, B_t, q_{95}), performance undergoes a continuous decline as the extrapolation distance increases. While the PFNN architecture supresses the rate of this decline, it does not eliminate the downward trend. For energy-related parameters (W_e, β_N), performance remains relatively stable within a certain operational range surrounding the training domain. However, beyond a critical extrapolation threshold, a sharp degradation occurs. The reason could be related to the transition of energy confinement mode when W_e and β_N are enhanced. This contrast might help future devices design a safer parameter ramp-up paths.

4. REAL-TIME IMPLEMENTATION

Building on the insights from previous analysis, the PFNN algorithm is deployed with carefully staged extrapolation steps from established operational regimes in HL-3. And it has enabled successful prediction and mitigation of most disruptions during HL-3's recent high-parameter campaigns. This section presents online disruption prediction and protection results from two key commissioning scenarios, high β_N operation and high I_p operation at 1.6 MA.

4.1 High beta disruption prediction and mitigation

HL-3's high β_N commissioning are focused on during the last experimental campaign, which starts from Shot 9285. Prior to this campaign, the maximum β_N in the available development dataset was 3.3. A disruption prediction algorithm trained on this pre-campaign data achieved an AUC of 0.977. During high β_N operation, the algorithm's real-time output is used to trigger the Massive Gas Injection (MGI) system for disruption mitigation.

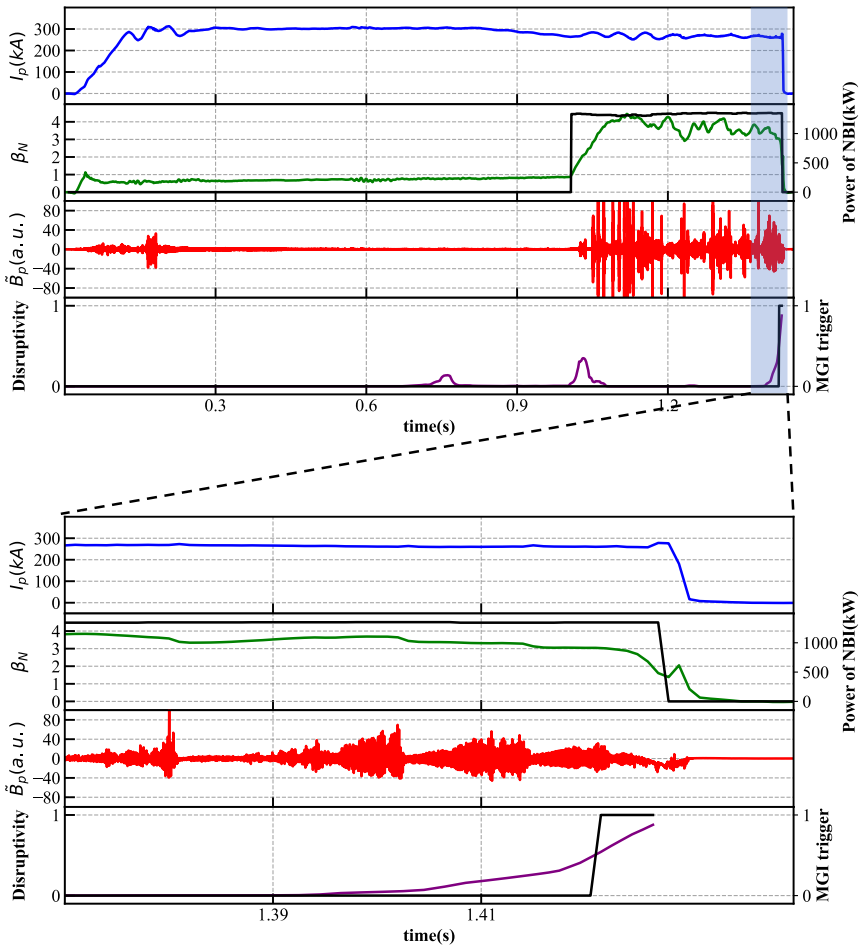

Fig 5 Demonstration of close-loop disruption prediction and mitigation in HL-3, during a high β_N discharge (shot 12478).

Figure 5 illustrates a representative experimental discharge. Following neutral beam injection (NBI) initiation at t=1.0 s, the plasma experiences significant perturbation due to the substantial momentum input from the NBI, resulting in plasma position oscillations. During this initial phase, the predicted disruptivity exhibits a modest increase, albeit of low amplitude. Subsequently, as the plasma configuration stabilizes, the β_N progressively increases with absorbed heating power, reaching values approaching 4. Concurrently, the β_N signal indicates the development of MHD instabilities, hindering further β_N increase. After t=1.4 s, the specific plasma pressure begins to decline, accompanied by further confinement deterioration. Consequently, the predicted disruption risk rises, triggering the MGI system and successfully mitigating the impending disruption.

Beyond this representative case, the system operated continuously during HL-3 discharges #12456 to #12509. Excluding invalid shots (e.g., failed discharges and plasma-less commissioning pulses), the closed-loop system executed in 22 valid discharges. Correct prediction output was achieved in 21 of these cases, both for disruptive and non-disruptive shots, yielding an overall reliability of 95.5%.

4.2 High current disruption prediction and soft-landing

The 1.6 MA plasma current operation on HL-3 was implemented during the final phase of its third experimental campaign, encompassing shots 6800 to 6985. Prior to this phase, the accumulated dataset comprised 2,163 shots, with a maximum plasma current of only 1.15 MA. A disruption prediction algorithm trained on this dataset achieved an AUC of 0.982 within its non-extrapolated operational regime and was deployed for real-time disruption warnings. During this phase, warning signals were not used to trigger MGI. Instead, a soft-landing strategy was implemented to rapidly ramp down the plasma current. This approach ensured minimal perturbation to wall condition between shots, thereby preserving the integrity of the commissioning schedule.

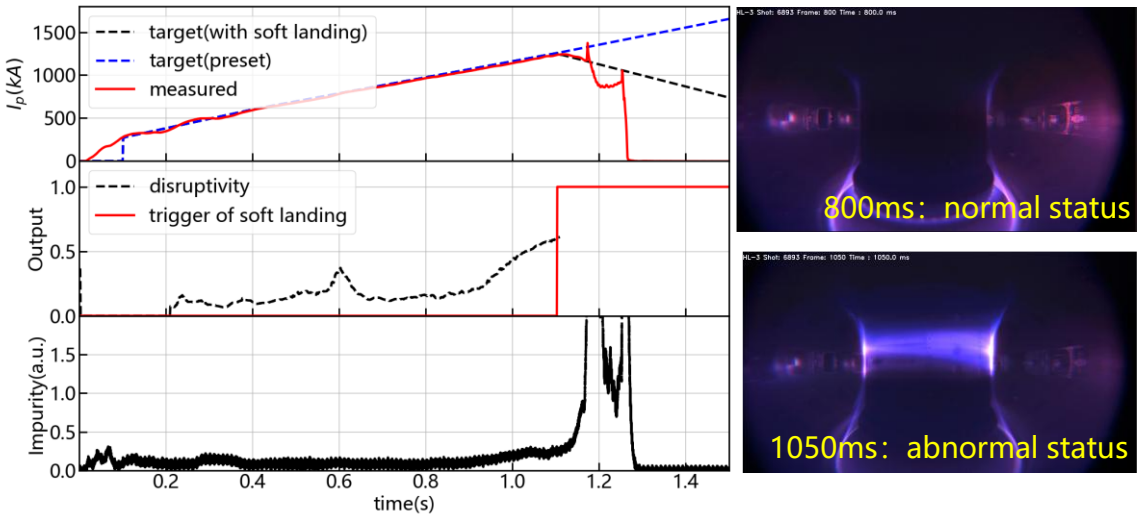

Fig 6 PFNN predicted the high-field side impurity source induced disruption and triggered soft landing control during Shot 6893

Figure 6 illustrates a closed-loop experimental shot where the algorithm successfully triggered soft landing control in response to a disruption precursor caused by impurity accumulation. Without intervention, the disruption would have occurred near 1.4 MA. The proactive mitigation reduced the disruption current to 0.9 MA, significantly reduced its severity.

5. SUMMARY

This study investigated the extrapolation capability of deep learning disruption prediction algorithms on HL-3, a device currently in its performance ramp-up phase. As expected, standard disruption prediction techniques exhibit significant performance degradation when extrapolating to new operational regimes, while PFNN effectively mitigates this degradation, enhancing extrapolation robustness. On the other hand, algorithm performance degrades differently when extrapolating along electromagnetic parameters versus energy-related parameters. This difference may inform the development of safer parameter escalation pathways for future fusion reactors.

Leveraging these insights, HL-3 successfully implemented closed-loop disruption prediction and mitigation for high β_N and high current scenarios. These demonstrated capabilities provide valuable operational experience expected to significantly inform the commissioning strategies of future fusion devices like ITER.

ACKNOWLEDGEMENTS

This work is supported by National MCF R&D program of China under Grant No.2024YFE03240100, National Natural Science Foundation of China under Grant No. U21A20440 and Natural Science Foundation of Sichuan Province under Grant No. 2024NSFSC1335. The authors wish to thank all the members at Southwestern Institute of Physics for doing their best to co-operate during the collection of dataset and development of algorithm.

REFERENCES

[1] ITER Physics Expert Group on Disruptions, Plasma Control, and MHD and ITER Physics Basis Editors 1999 Nuclear Fusion 39 2251.

[2] H. Jin, Q.M. Hu, N.C. Wang, et al. 2015 Plasma Physics and Controlled Fusion 57: 104007.

[3] M. Lehnen, K. Aleynikova, P.B. Aleynikov, et al. 2015 Journal of Nuclear Materials 463: 39-48.

[4] P.C. de Vries, G. Pautasso, D.A. Humphreys et al. et al. Fusion Science and Technology 69:471-484.

[5] J.K. Harbeck, A. Svyatkovskiy and W. Tang 2019 Nature 568: 526–531.

[6] C. Rea, K.J. Montes, K.G. Erickson, et al. 2019 Nuclear Fusion 59: 096016.

[7] B.H. Guo, D.L. Chen, C. Rea, et al. 2023 Nuclear Fusion 63: 094001.

[8] Z.Y. Yang, F. Xia, X.M. Song, et al. 2020 Nuclear Fusion 60: 016017.

[9] J. Lee, S. H. Hahn, H. Han, et al. 2025 Nuclear Fusion 65: 056040

[10] W. Zheng, F. Xue, Z.Y. Chen, et al. 2023 Communications Physics 6: 181

[11] Y. Zhong, W. Zheng, Z. Y. Chen, et al. 2021 Plasma Physics and Controlled Fusion 63 075008.

[12] C. Shen, W. Zheng, Y. Ding, et al. 2023 Nuclear Fusion 63:046024

[13] X.R. Duan, M. Xu, W.L. Zhong, et al. 2022 Nuclear Fusion 62: 042020

[14] X.R. Duan, M. Xu, W.L. Zhong, et al. 2024 Nuclear Fusion 64: 112021

[15] Z.Y. Yang, W.L. Zhong, F. Xia, et al. 2025 Nuclear Fusion 65: 026030

[16] A. Murari, R. Rossi, E. Peluso, et al. 2020 Nuclear Fusion 60: 056003.

[17] J.X. Zhu, C. Rea, R.S. Granetz, et al. 2023 Nuclear Fusion 63: 046009.

[18] N. N. Wu, Z. Y. Yang, R. P. Li, et al. Preprint at https://doi.org/10.48550/arXiv.2409.09238