IN-DEPTH RESEARCH ON THE INTERPRETABLE DISRUPTION PREDICTOR IN HL-2A

Zongyu Yang[1&2], Fan Xia[1] , Xianming Song[1] , Zhe Gao[2] , Shuo Wang[1] , Yunbo Dong[1]

[1] Southwestern Institute of Physics, P.O. Box 432, Chengdu 610041, China
[2] Department of Engineering Physics, Tsinghua University, Beijing, 100084, China

Introduction

A series of in-depth researches are implemented on the disruption predictor in HL-2A, mainly for 2 aims, accuracy and interpretability.

➢ **For further improvement of accuracy**

- 4 adjustments are tried to solve 4 corresponding problems in the baseline model. These optimizations increase the model's AUC (Area Under receiver operating characteristic Curve) from 0.905 to 0.944.

➢ **For interpretability of model**

- An interpretation method is proposed to evaluate the importance of each input signal when deciding the model's output. The result of single shot interpretation shows good coherence with the causes of disruption.
- Shot Nos.20000-36000 are manually analyzed to make a disruption cause dataset. Statistical analysis of the interpretation algorithm' output on this dataset also shows a good coherence with the disruption causes.
- A Bayes classifier is developed to recognize the cause of disruption based on the interpretation algorithm's output. This classifier has an accuracy of 71.2% on the labelled dataset, which contains 5 disruption causes. and 605 disruptive shots.

Disruption prediction dataset

➢ Shot count for training and validation:
Shot Nos. 20000-33000 in HL-2A, 2800 non-disruptive shots and 1005 disruptive shots
➢ Shot count for testing:
Shot Nos. 33000-36000 in HL-2A, 816 non-disruptive shots and 290 disruptive shots
➢ Input signal list:
As shown in the table
➢ Pre-processing method:

$$X_{norm} = \begin{cases} 10, & \dfrac{X-\bar{X}}{D} \geq 10 \\ \dfrac{X-\bar{X}}{D}, & 10 < \dfrac{X-\bar{X}}{D} < 10 \\ -10, & \dfrac{X-\bar{X}}{D} \leq 10 \end{cases}$$

➢ Other details are same with our previous research *[Zongyu Yang et al 2020 Nucl. Fusion 60 016017]*.

Signal name	Sample rate(kHz)	Physical meanings
I_p	1	plasma current
Target_I_p	1	target plasma current
V_Loop	1	loop voltage
B_t	1	toroidal magnetic field
I_Ohm	1	current in Ohmic field coil
Bolometer	1	power of radiation measured by bolometer
density	1	density of electrons at the centre of plasma
HardX_1	1	power of hard-x-ray (0–5 MeV)
HardX_2	1	power of hard-x-ray (5–10 MeV)
P_ECRH	1	Power of ECRH
P_NBI	1	Power of NBI
EFIT_q_bdry	1	safety factor at the boundary of plasma calculated by EFIT
EFIT_r	1	minor radius of plasma calculated by EFIT
EFIT_R	1	position of geometric centre in the radial direction calculated by EFIT
EFIT_Z	1	position of geometric centre in the vertical direction calculated by EFIT
EFIT_li	1	internal inductance calculated by EFIT
StoredEnergy	1	energy stored in plasma
betaN	1	normalized beta
Da_divertor	10	D-α ray at divertor
SoftX	10	power of soft-x-ray
Mirnov_Tor_A/B	10	a pair of toroidal probes located at symmetric positions
Mirnov_Pol_A/B	10	a pair of poloidal probes located at symmetric position

Baseline model

Although a previous model with high accuracy has been proposed in our previous research. The previous model has about 4 million parameters, thus it is hard to realize real-time prediction. Therefore a new version of model is proposed to be the baseline model. Structure of the baseline model is shown in figure 1. Performance of the model is listed here.

➢ Number of parameters: 0.1 million ☺
➢ Time cost of each input slice: 2ms ☺
➢ Accuracies:
TPR0.832/TNR0.825/AUC0.905☹

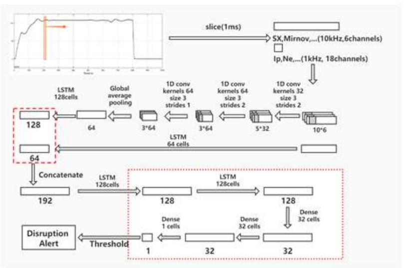

Figure 1 Structure of baseline model

Optimization methods and Comparison experiments

Challenges	Solutions
Multimodal data The input signals come from different sources and have different characteristics	**1.5-D structure** Signals from different sources are dealt separately at first and merged in the middle layer of model
Variable precursor time Different types of disruptions have very different precursor time.	**Fuzzy labels in disruptive shots** TTD < 30ms→1 TTD > 200ms→0 30ms < TTD < 200ms→ None
Auxiliary heating The switch on/off of auxiliary heating brings a sudden change of environment, which means the criterion of algorithm should change, too.	**Preset control signal** The control signals of auxiliary heating are set before experiment. Thus they can be put into to the algorithm in advance.
Time variance of device The situation of diagnostic system and control system in device varies with time	**Fine-tune on latest data** Use complete dataset to train the model firstly. Then fine-tune the top layers (dotted box in Figure 1) of model on the latest part of dataset.

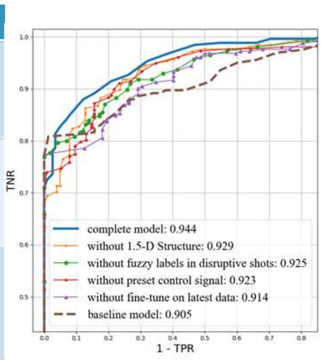

Figure 2 ROC Curve of each version of the model. TPR (True Positive Rate) means the proportion of disruptive shots that are correctly predicted. TNR (True Negative Rate) means the proportion of non-disruptive shots that are correctly predicted. Numbers in the legends are AUC of each model

Model interpretation method

➢ Add random noises to each input signal of model in turn, and the respective change in final output of model would indicate the importance of corresponding signal.
➢ Note that the noise should be added to the middle layer output in the model, which is indicated with dashed box in figure 1. There are 3 reasons:
✓ middle layer output in neural networks with batch-normalization methods tends to be in a gaussian distribution, therefore the random noise will cause similar effect on each input signal.
✓ Data from all input signals are still individual at this location. So noise can added to each signal separately.
✓ Only the top layers need to be rerun for 24 times, which greatly reduces of the computational expense

Interpretable model: single shot interpretation

Lock-mode induced disruption: 35104
✓ Most related signals: Mirnov probe signals

Density limit induced disruption: 35240
✓ Most related signal: density

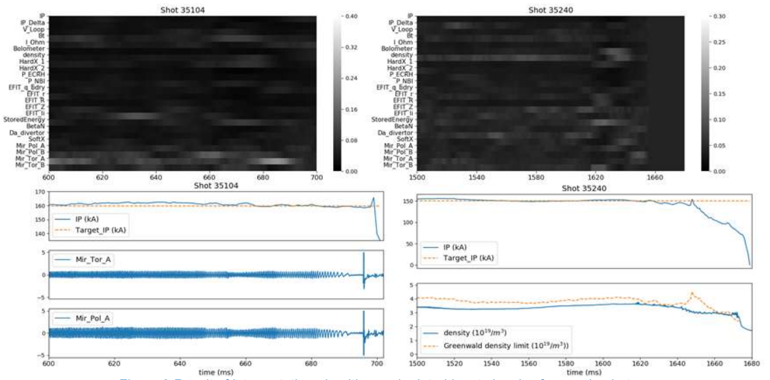

Figure 3 Result of interpretation algorithm and related input signals of example shots.

Disruption cause dataset

➢ Shot Nos. 20000-36000 are analyzed manually to find the causes of disruptions in HL-2A. Among them 613 shots with clear causes are selected to make up a disruption cause dataset.
➢ The researches in the two subsequent sections are implemented on this dataset

Type name	Shot count
horizontal displacement	67
vertical displacement	55
lock mode	253
radiation	170
low q boundary	8
density limit	60

Interpretable model: statistical analysis

➢ As expected, the most important signals for vertical displacement, lock mode, radiation, low q on boundary and density limit induced disruptions are EFIT_Z, Mirnov probe, Bolometer/SoftX, EFIT_q_bdry and density, respectively.
➢ The result of horizontal displacement seems to be kind of complex. It is suspected that other causes might also result in horizontal displacement, which calls for a further investigation.
➢ The heights of bars come from the equation below. Here $\bar{\mathrm{I}}(c,s)$ means the averaged importance of signal *s* in cause *c* induced disruptions. $\mathrm{I}(i,s)$ means the importance of signal *s* in shot No. *i*. $\sum_{i\ in\ c}\mathrm{I}(i,s)$ means the sum of $\mathrm{I}(i,s)$ on all the shots with the cause of *c*. $\sum_i \mathrm{I}(i,s)$ means the sum of $\mathrm{I}(i,s)$ on all the shots in dataset. The $\sum_i 1$ and $\sum_{i\ in\ c} 1$ are counts of shots used to calculate the mean value.

$$\bar{\mathrm{I}}(c,s) = \frac{\sum_{i\ in\ c}\mathrm{I}(i,s)}{\sum_{i\ in\ c} 1} - \frac{\sum_i \mathrm{I}(i,s)}{\sum_i 1}$$

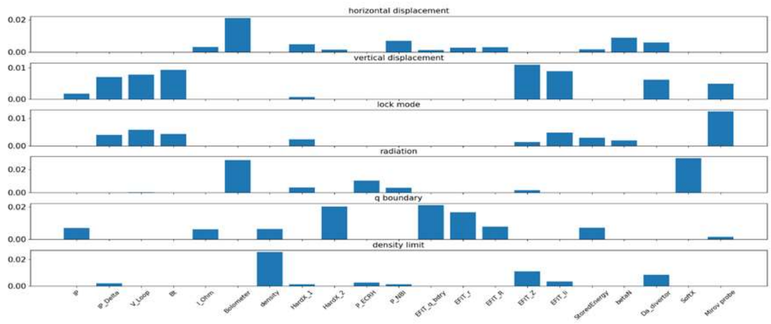

Figure 4 Averaged importance of each input signal among shots of each disruption cause.

Interpretable model: disruption cause recognizer

Dataset
➢ The type of low safety factor in boundary has only 8 shots and therefore are abandoned.
➢ Finally 605 shots are reserved.

Model
➢ Considering the limited size of the dataset, a naive Bayes model based on gaussian kernel function is selected to be the classifier.
➢ Input: the importance of each input signal averaged among the time range before the alarm triggered by disruption prediction algorithm, i.e. a vector of 24 elements.
➢ Output: cause of disruption

Result
➢ 10-fold cross validation
➢ Top-1 accuracy: 71.2%(431/605).

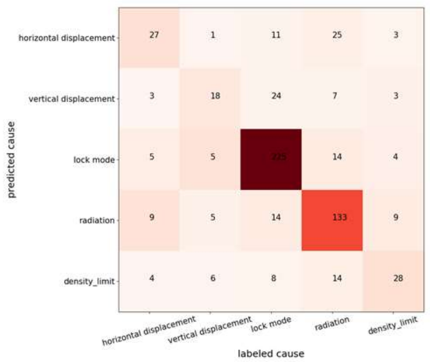

Figure 5 Confusion matrix of disruption cause recognizer on dataset by 10-fold cross validation

Future works

➢ Online testing of the disruption predictor.
➢ Validating existing algorithms and experiences on other tokamaks.
➢ Further investigation is still needed on how to reach a high accuracy with a limited computational expense

***Email:* zy-yang@swip.ac.cn**

中核集团 CNNC 核工业西南物理研究院 Southwestern Institute of Physics