

Recent progress on deep learning-based disruption prediction algorithm in HL-2A tokamak

Zongyu Yang(杨宗谕), Yuhang Liu(刘宇航), Xiaobo Zhu(朱晓博), Zhengwei Chen(陈正威), Fan Xia(夏凡), Wulyu Zhong(钟武律), Zhe Gao(高喆), Yipo Zhang(张轶泼), and Yi Liu(刘仪)

Citation: Chin. Phys. B, 2023, 32 (**7**): 075202. DOI: 10.1088/1674-1056/accb44
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Disruption prediction based on fusion feature extractor on J-TEXT

Wei Zheng(郑玮), Fengming Xue(薛凤鸣), Zhongyong Chen(陈忠勇), Chengshuo Shen(沈呈硕), Xinkun Ai(艾鑫坤), Yu Zhong(钟昱), Nengchao Wang(王能超), Ming Zhang(张明),Yonghua Ding(丁永华), Zhipeng Chen(陈志鹏), Zhoujun Yang(杨州军), and Yuan Pan(潘垣)

Chin. Phys. B, 2023, 32 (**7**): 075203. DOI: 10.1088/1674-1056/acc7fc

The intermittent excitation of geodesic acoustic mode by resonant Instanton of electron drift wave envelope in L-mode discharge near tokamak edge

Zhao-Yang Liu(刘朝阳), Yang-Zhong Zhang(章扬忠), Swadesh Mitter Mahajan, A-Di Liu(刘阿娣), Chu Zhou(周楚), and Tao Xie(谢涛)

Chin. Phys. B, 2022, 31 (**4**): 045202. DOI: 10.1088/1674-1056/ac43ac

A divertor plasma configuration design method for tokamaks

Yong Guo(郭勇), Bing-Jia Xiao(肖炳甲), Lei Liu(刘磊), Fei Yang(杨飞), Yuehang Wang(汪悦航), Qinglai Qiu (仇庆来)

Chin. Phys. B, 2016, 25 (**11**): 115201. DOI: 10.1088/1674-1056/25/11/115201

Toroidicity and shape dependence of peeling mode growth rates in axisymmetric toroidal plasmas

Shi Bing-Ren (石秉仁)

Chin. Phys. B, 2014, 23 (**7**): 075206. DOI: 10.1088/1674-1056/23/7/075206

Numerical study of physical properties of resistive wall modes in tokamaks

Xia Xin-Nian (夏新念), Liu Yue (刘悦), Liu Chao (刘超), He Yu-Ling (何玉玲), Xia Guo-Liang (夏国良)

Chin. Phys. B, 2013, 22 (**5**): 055203. DOI: 10.1088/1674-1056/22/5/055203

TOPICAL REVIEW — Plasma disruption

Recent progress on deep learning-based disruption prediction algorithm in HL-2A tokamak

Zongyu Yang(杨宗谕)[1,†], Yuhang Liu(刘宇航)[1], Xiaobo Zhu(朱晓博)[1], Zhengwei Chen(陈正威)[1], Fan Xia(夏凡)[1], Wulyu Zhong(钟武律)[1,‡], Zhe Gao(高喆)[2], Yipo Zhang(张轶泼)[1], and Yi Liu(刘仪)[1]

[1] *Southwestern Institute of Physics, Chengdu* 610043, *China*
[2] *Tsinghua University, Beijing* 100084, *China*

(Received 18 January 2023; revised manuscript received 3 April 2023; accepted manuscript online 7 April 2023)

Disruption prediction and mitigation is a crucial topic, especially for future large-scale tokamaks, due to disruption's concomitant harmful effects on the devices. On this topic, disruption prediction algorithm takes the responsibility to give accurate trigger signal in advance of disruptions, therefore the disruption mitigation system can effectively alleviate the harmful effects. In the past 5 years, a deep learning-based algorithm is developed in HL-2A tokamak. It reaches a true positive rate of 92.2%, a false positive rate of 2.5% and a total accuracy of 96.1%. Further research is implemented on the basis of this algorithm to solve three key problems, i.e., the algorithm's interpretability, real-time capability and transferability. For the interpretability, HL-2A's algorithm gives saliency maps indicating the correlation between the algorithm's input and output by perturbation analysis. The distribution of correlations shows good coherence with the disruption causes. For the transferability, a preliminary disruption predictor is successfully developed in HL-2M, a newly built tokamak in China. Although only 44 shots are used as the training set of this algorithm, it gives reasonable outputs with the help of data from HL-2A and J-TEXT. For the real-time capacity, the algorithm is accelerated to deal with an input slice within 0.3 ms with the help of some adjustments on it and TFLite framework. It is also implemented into the plasma control system and gets an accuracy of 89.0% during online test. This paper gives a global perspective on these results and discusses the possible pathways to make HL-2A's algorithm a more comprehensive solution for future tokamaks.

Keywords: macroinstabilities, tokamaks, neural networks, magnetic confinement and equilibrium

PACS: 52.35.Py, 52.55.Fa, 07.05.Mh, 52.55.–s **DOI:** 10.1088/1674-1056/accb44

1. Introduction

Disruption places the tokamak-based fusion power plant a dilemma. The ignition condition and economical efficiency call for large-scale devices and high parameter plasma,[1] while in these conditions the risk of thermal load, electromagnetic force and runaway electron beam during disruptions could be unacceptable.[2] Disruption prediction and mitigation techniques take the responsibility to resolve this conflict. By accurately predicting the disruptions and mitigate them in advance, their harmful effects can be distinctly alleviated.[3,4]

In experiment, disruptions could be induced by various of causes and event chains, e.g., mode locking, radiation collapse and MHD instabilities.[5] The complexity of their precursors makes the prediction a tough task. There are mainly 2 pathways to solve the problem. One is to design a dedicated framework to monitor disruption-related events and analysis the evolution of event chains, which leads to some representative programs like DECAF (disruption event characterization and forecasting).[6] The other way is to use data-driven algorithms, such as deep learning, to build an end-to-end model by learning from labeled disruptive and non-disruptive data. In recent years, many data-driven algorithms have been developed on various tokamaks, such as DIII-D, JET, C-Mod, J-TEXT and EAST.[7–14]

The data-driven method has been tried in HL-2A at around 2010. A multi-layer perceptron model is built to predict the disruptions according to the evolution of signals from bolometer array.[15] However, the ability of machine learning to comprehensive analysis multi-diagnostic signals is not very mature then. The performance of algorithm is limited since only a part of the disruptions will be accompanied with precursors on bolometer array.

Since 2018, deep learning method is tried in HL-2A. A disruption prediction algorithm based on 1.5 dimensional convolutional neural network and long short-term memory recurrent neural network (1.5D CNN+LSTM) is proposed and shows great performance by combining information from 25 disruption-related signals.[16] Therefore, a series of subsequent researches are implemented around this algorithm, such as the interpretation of algorithm,[17] real-time implementation in plasma control system (PCS),[18] cross-tokamak prediction and data-physics dual-driver algorithm.[19] These researches aim to make the disruption prediction algorithm a reliable, regular operated and operator friendly module in the

†Corresponding author. E-mail: author:zy-yang@swip.ac.cn
‡Corresponding author. E-mail: zhongwl@swip.ac.cn

© 2023 Chinese Physical Society and IOP Publishing Ltd

PCS and provide a possible solution for future large-scale tokamaks. In this paper, we give an overview on the progresses about deep learning-based algorithm in HL-2A and discusses the possible pathways to further promote the algorithm.

The rest part consists of 5 sections. Section 2 introduces the backbone of HL-2A's disruption prediction algorithm. Then the backbone model is modified or combined with supplement modules to solve three key problems for disruption prediction in future fusion reactors, i.e., interpretability, real-time capacity and transferability. In Section 3, an integrated disruption prediction and mitigation system is developed in HL-2A. It realizes the real-time prediction and mitigation during online experiment, which could be a solution for the problem of real-time capacity. Section 4 gives the method and main results of model interpretation, which answers the question of interpretability. Section 5 focuses on the issue of transferability, i.e., how to develop an algorithm with limited data on future tokamaks. And three possible solutions will be introduced with preliminary implementation examples. Finally, a summary and outlook is given in Section 6.

2. Disruption prediction backbone: 1.5D CNN + LSTM

Disruption prediction task can also be seen as a pattern recognition of disruption precursors, which is a classical problem for machine learning techniques. Previous machine learning-based disruption prediction algorithms often use vectors consist of scalar features as their input.[7,8,10] Therefore they are not so good at dealing with high dimensional, temporal and multi-diagnostic signals, which should contain much more information. This problem is solved by deep learning recently. Fusion recurrent neural network (FRNN) introduces the profile data to promote its performance.[13] And the hybrid neural network in DIII-D proves that temporal information is pretty helpful for disruption prediction task.[14] In HL-2A, a disruption predictor based on 1.5D CNN+LSTM is developed, which takes signals from lots of different diagnostics with possibly different sample rates. Dedicated structure is designed to solve the numerical instability when training with data from different sources. Finally the algorithm gets an accuracy of 96.1% on testing set, which mainly benefits from the wide range of input signals.[16]

In this part, Subsection 2.1 introduces the HL-2A disruption prediction dataset. Subsection 2.2 presents the data pipeline and neural network structure of HL-2A's algorithm. Subsection 2.3 shows the accuracy of algorithm and discusses the design philosophy of 1.5D CNN+LSTM.

2.1. Dataset

Three aspects are selected here to give a global presentation of HL-2A's disruption prediction dataset, i.e., shot numbers, input signal lists and typical disruption shots.

Table 1 shows the shot numbers used for the training, validation and testing of neural network. Almost all the shots after 2013 are used without any intended selection, unless some of the input signals are lacked or obviously abnormal. Therefore the dataset consists of many types of disruptions that ever appeared in HL-2A, such as density limit, mode locking, and low safety factor on boundary.

Table 1. Shot numbers used for training, validation, and testing of the HL-2A's disruption predictor.

	Disruptive shot	Non-disruptive shot
Training	1134	3157
Validation	284	721
Testing	475	1271

Table 2 gives the main information of the algorithm's 25 input signals. These signals are selected to cover the basic plasma parameters and disruption precursors as completely as possible. An important fact is that the signals have different statistical distributions and sample rates. This leads to a series of pre-processing methods and special neural network structures, which will be introduced in the subsequent part.

Table 2. Signals in HL-2A used as the input of disruption predictor. Their sample rate, statistical features and physical meanings are also listed here.[16]

Channel	Sample rate (kHz)	Mean/Std	Physical meanings	Reason of selection
$I_{\mathrm{plasma,target}}$	1	146/20	Target plasma current	Basic plasma parameter
I_{plasma}	1	145.6/19.5	Plasma current	
$\Delta I_{\mathrm{plasma}}$	1	–0.04/0.03	Relative difference between I_{plasma} and $I_{\mathrm{plasma,target}}$	Correlated to plasma current limit
B_{t}	1	1.34/0.004	Toroidal magnetic field	
$q_{\mathrm{a,EFIT}}$	1	4.41/0.82	Safety factor at plasma boundary given by EFIT	
P_{ECRH}	1	0.09/0.25	Power of ECRH heating	The abnormal or inappropriate heating power could introduce disruptions
P_{NBI}	1	0.12/0.25	Power of NBI heating	
$I_{\mathrm{Ohm_coil}}$	1	–0.005/6.84	Current in Ohmic field coil	
$\mathrm{GasPuff_{FB}}$	1	0.084/0.52	Gas injection (feedback)	To monitor unscheduled density increasing or decreasing
$\mathrm{GasPuff_{FF}}$	1	22/21	Gas injection (feedforward)	
$D_{\alpha,\mathrm{div}}$	10	0.4/0.15	D-α ray at divortor	To distinguish L/H mode and divertor/limiter plasma

Table 2. (*Continued*).

Channel	Sample rate (kHz)	Mean/Std	Physical meanings	Reason of selection
SoftX	100	4.38/3.69	Soft-x ray	Correlated to large sawtooth induced disruptions
r_{EFIT}	1	0.376/0.044	Minor radius given by EFIT	The loss of control on plasma position and shape could lead to disruptions.
R_{EFIT}	1	1.654/0.02	Radial position of geometric center given by EFIT	
Z_{EFIT}	1	0.002/0.02	Vertical position of geometric center in the EFIT	
W_{e}	10	7.85/10.22	Stored energy	The abnormal increasing or decreasing of stored energy could lead to disruptions
$N_{\mathrm{e,fir}}$	1	1.13/0.78	Electronic density	Correlated to density limit
$n_{\mathrm{Greenwald}}$	1	3.34/0.61	Greenwald density limit	
$P_{\mathrm{radiation}}$	1	168.7/40.4	Power of radiation measured by bolometers	The abnormal increasing of plasma radiation power could lead to disruptions
$\mathrm{Mir}_{\mathrm{Pol04}}$	100	0.0078/0.226	A pair of poloidal probes located at the symmetric position	Correlated to mode locking induced disruptions
$\mathrm{Mir}_{\mathrm{Pol13}}$	100	0.0025/0.226		
$\mathrm{Mir}_{\mathrm{Tor04}}$	100	–0.0039/0.113	A pair of toroidal probes located at the symmetric position	
$\mathrm{Mir}_{\mathrm{Tor09}}$	100	–0.0008/0.113		
$\mathrm{HX}_{0\sim5\ \mathrm{MeV}}$	100	0.0044/0.02	Power of hard-x ray(0–5M eV)	To monitor runaway electrons, which might lead the loss of control on plasma current
$\mathrm{HX}_{0\sim10\ \mathrm{MeV}}$	100	0.02/0.18	Power of hard-x ray (0–10 MeV)	

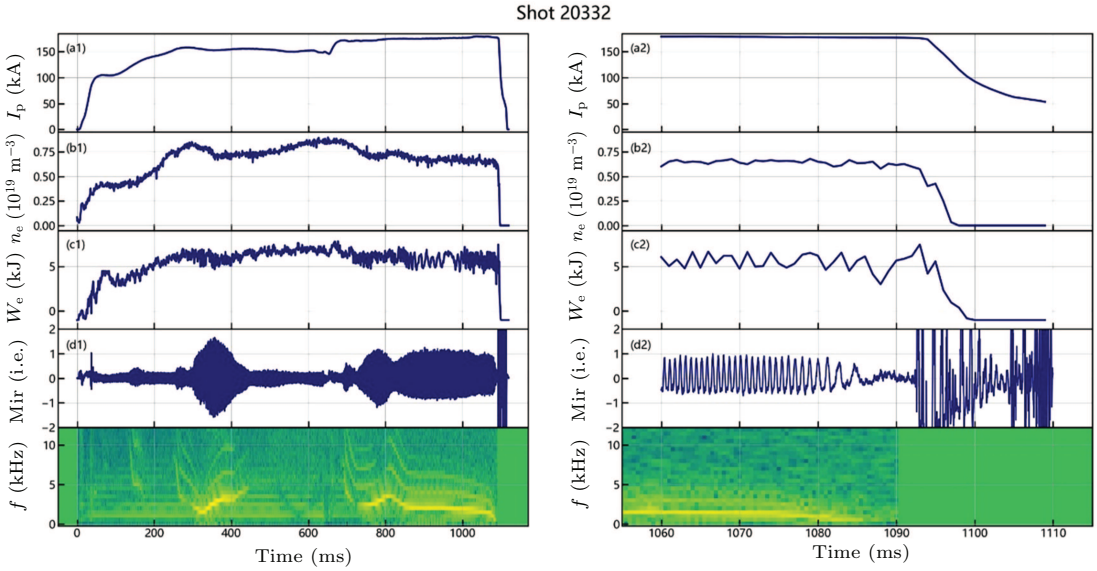


Fig. 1. The evolution of main plasma parameters, (a) plasma current, (b) line averaged electron density, (c) plasma stored energy measured by diamagnetic method, (d) magnetic perturbation measured by Mirnov probe, (e) spectrogram of the magnetic perturbation signal, during a $m/n = 2/1$ mode locking-induced disruptive shot in HL-2A. The left subfigures plot the evolution during the whole shot. And the right subfigures plot the evolution close to disruption. A mode-locking process can be found on the Mirnov probe signal in the right bottom subfigure, which takes place at about 15 ms before the disruption.

Figure 1 presents a mode locking-induced disruptive shot in HL-2A. The plasma current in HL-2A is normally 150 kA and the line averaged electron density is normally 0.5×10^{19}–3×10^{19} m^{-3}. The time scale of mode locking process in HL-2A is around 15 ms. During disruptions, the slope of plasma current is around 10 MA/s, which leads to a current quench time of 5–20 ms. Many types of disruption precursors can be found in HL-2A's experimental data, including but not limited to mode-locking, density limit, plasma displacement, radiation collapse and low safety factor. The variety and short timescale of disruption precursors make the disruption prediction a tough task.

2.2. Data pipeline and neural network structure

The working cycle of HL-2A's disruption predictor is determined as 1 ms by the PCS. During each cycle, the algorithm takes the signals in past 20 ms from the channels given in Table 2. After that, pre-processing is taken to make these signals into structured and normalized matrix. Then the neural network takes the matrix as input and returns a possibility of disruption as output. Finally the PCS decides if disruption mitigation techniques should be triggered according to the output.

The pre-processing consists of 3 steps.

(a) Resampling all the signals into 100 kHz.

(b) Reducing data from each channel in Table 2 by the channel's mean value, and dividing them by standard deviation.

(c) Replacing the values larger than 10 with 10 and values smaller than -10 with -10.

After pre-processing, the input signal for each cycle will be uniformed into a 25 channels, 20 ms long and 100 kHz matrix, whose size should be 2000×25.

The structure of neural network is presented in Fig. 2, which consists of convolutional layers, LSTM layers and fully connected layers. The cell number of the last fully connected layer is 2, therefore the neural network will output 2 values for each cycle, representing the possibility of disruption and non-disruption.

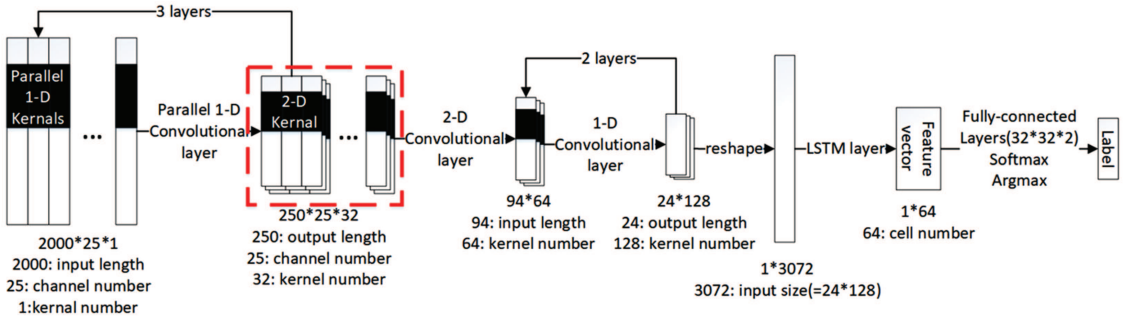


Fig. 2. Structure of 1.5D CNN+LSTM. This figure is adapted from Fig. 3 in Ref. [16], where only the 1.5D CNN is presented. LSTM is merged into the figure after adaptation.

2.3. The performance of 1.5D CNN + LSTM

The performance of disruption prediction algorithm is often evaluated by two indicators, i.e., true positive rate (TPR) and true negative rate (FPR). The former refers to the possibility that the algorithm accurately gives an alarm during a disruptive shot. And the latter refers to the possibility that the algorithm gives a false alarm during a non-disruptive shot.

In this research, the deep learning disruption predictor has a TPR of 92.2% and a FPR of 2.5%, which is a pretty good result. A key promotion is brought by its design to separately deal with signals from different diagnostics at first and merge them in the middle layer of neural network. The structure is more suitable for fusion data, which is an array consists of many heterogeneous parameters with various statistical distributions and waveforms, rather than a structured matrix like images or sentences. Table 3 gives the result of ablation experiment for this design, where 2D CNN means to merge signals from different diagnostics at the first layer of neural network, and 1D CNN means to deal with signals separately all the time. It is obvious that the 1.5D CNN gets the best performance.

Table 3. Shot numbers used for training, validation, and testing of the HL-2A's disruption predictor.

	TPR (%)	FPR (%)
1.5D CNN+LSTM	92.2	2.5
1D CNN+LSTM	70.7	5.8
2D CNN+LSTM	74.6	9.0

3. Real-time disruption prediction and mitigation in HL-2A

To implement the disruption prediction algorithm in the PCS, two main adjustments should be made on the version described in Section 2. Firstly, the principle to select input signals in Section 2 is to cover the disruption precursors as completely as possible. But not all these signals can be acquired in HL-2A's PCS due to some engineering limitation of data acquisition system (DAS). Therefore the input signal list will be kind of different from the one in Table 2, which might be harmful to the algorithm's performance. Secondly, some adjustments on neural network structure are required to accelerate its calculation speed into 1 ms per cycle. The detailed description about these updates can be found in Ref. [17].

After that, the integrated disruption prediction and mitigation system can be built based on the real-time algorithm. Figure 3 shows the framework of this integrated system. The DAS gathers all the needed diagnostic signals and sends them to the PCS. These data are firstly used to do the position and shape control, in which process some secondary signals are produced. Then these secondary signals, together with the raw diagnostic signals, are sent to the disruption prediction module. In HL-2A, the PCS is mainly developed in C language while the disruption prediction algorithm is developed in Python, so a cross language interaction is required. Thus, there is a C-based disruption prediction module, which organizes the input data and calls the Python-based disruption prediction module. Finally, the prediction result is sent back to the C-based module to decide whether or not a trigger signal should be sent to the disruption mitigation system.

To validate the reliability of the integrated system, open-loop online testing is tried in 382 shots during shot Nos. 38650–39347. A TPR of 95.8% and a FPR of 22.5% are got during the online testing, which is acceptable but not as good as the result in Subsection 2.3 due to the limited input signal list and the unpredictable interference in DAS.

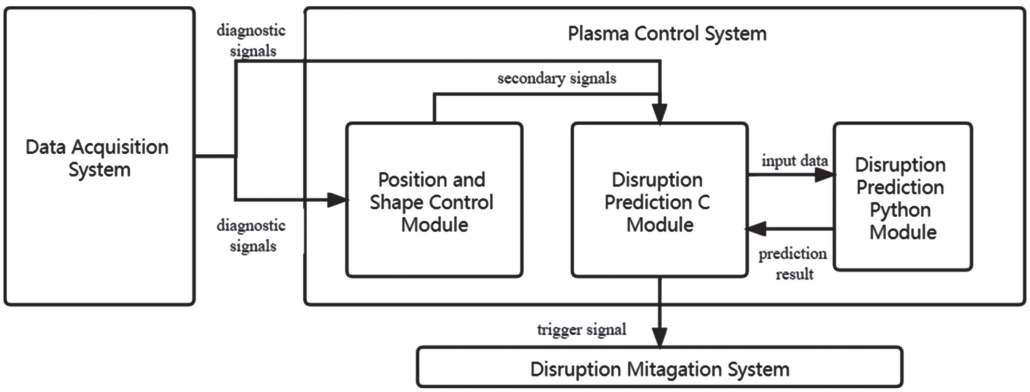


Fig. 3. Design of the real-time disruption prediction and mitigation system.[18]

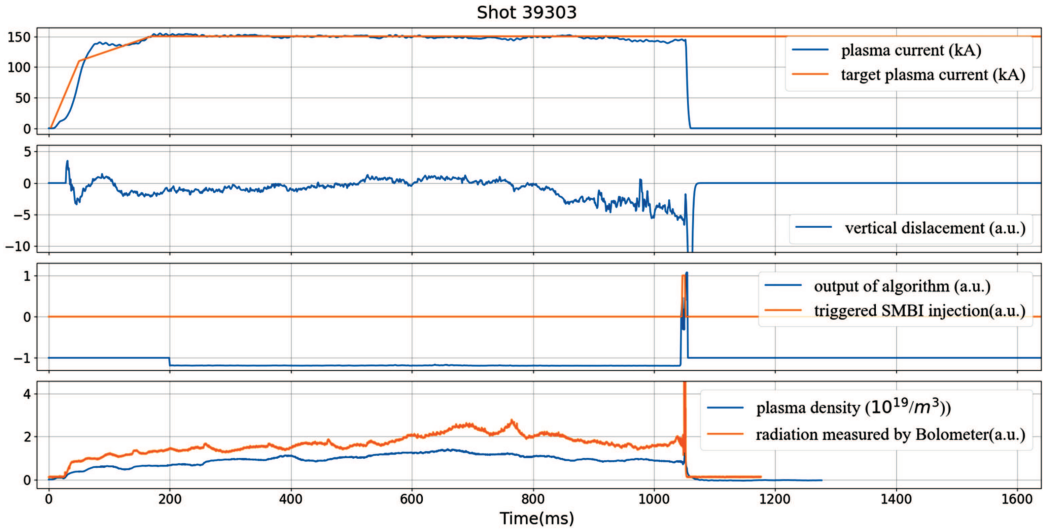


Fig. 4. A demo shot of real-time disruption prediction and mitigation. A vertical displacement takes place after 800 ms due to the improper control parameters of the upper divertor coil, as indicated in the second subfigure. The output of algorithm increases rapidly when the vertical displacement amplitude becomes too large. Then the SMBI system is triggered to implement the disruption mitigation, as shown in the third subfigure. Finally a mitigated disruption occurs. Rapid decline can be seen on signals of plasma current, plasma density and bolometer radiation.[18]

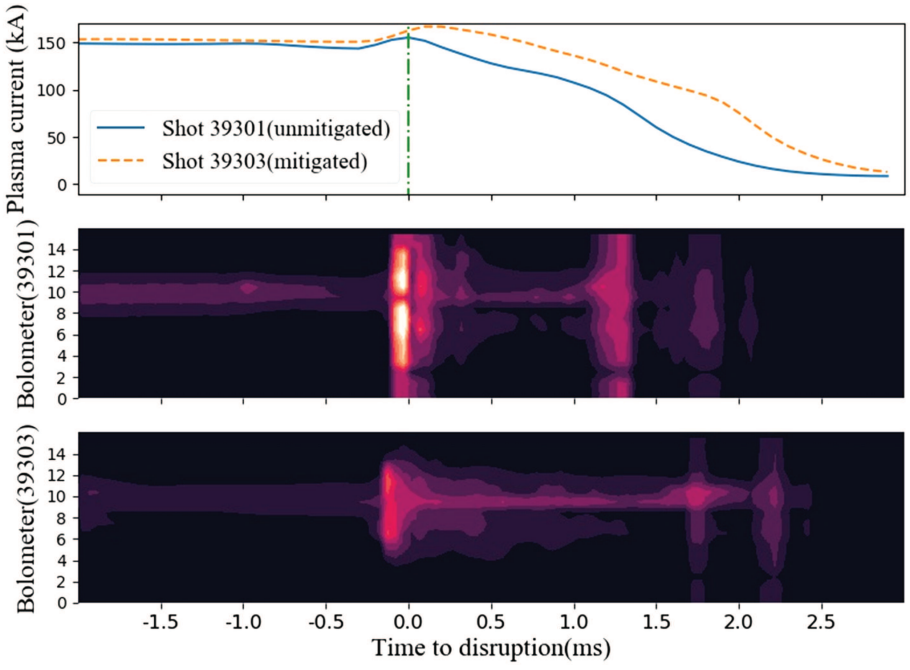


Fig. 5. Comparison on the current quench and thermal quench processes between shot 39303 (mitigated) and shot 39301 (unmitigated). The thermal quench process is observed by upper bolometer array with 16 channels.[18]

After the system validation, closed-loop testing is implemented. Figure 4 presents a demo shot, i.e., shot 39303, where a vertical displacement-induced disruption is predicted and then mitigated by SMBI, showing the system's ability to real-timely predict and mitigate the coming disruptions. Figure 5 compares the thermal quench and current quench process between shot 39303 and an unmitigated disruption, shot 39301. During shot 39303, the thermal energy of plasma is released by radiation within a longer time range and the peak value is clearly lower than in the unmitigated one.[18]

4. Prediction with interpretability and cause recognition

In most situations, deep learning is considered as an end-to-end black-box algorithm, which takes the disruption related signals as input and returns a probability of disruption without any explanation. Such an opaque system is difficult to accept for a fusion reactor with high safety requirements. In this section, an interpretation method is proposed to crack the black box. Subsection 4.1 gives the detail of the interpretation method. Subsection 4.2 shows the result of interpretation on single shot or from statistical point of view. Subsection 4.3 describes a preliminary disruption reason recognizer developed on top of the interpretation method. These results show that deep learning has the potential to realize high accuracy and interpretability at the same time.[17]

4.1. Interpretation method

The interpretation method is based on the 1.5-D CNN structure. In this structure a special node exists, as marked with red box in Fig. 2, where the input signals have been dealt with by several neural network layers and will be concatenated in the next layer. On this node, a Gaussian noise could be added on data from each input channel to eliminate their contribution to the output. Then a distribution of correlation between each input channel and the algorithm's output could be generated by perturbing the channels in turn and recording their contributions.

4.2. Interpretation results

Figure 6 gives the result of interpretation on 2 example shots. The 2 heatmaps are temporal evolution of correlation between algorithm's output and each input channel. The brighter color means the corresponding input channel on y-axis is more related to disruption. In shot 35104, an obvious mode-locking phenomenon can be observed in the Mirnov probe signals before the disruption, as shown in the left bottom subfigures. And the rows in the left heatmap corresponding to the Mirnov probes are brighter as expected. In shot 35240, the plasma density gradually increases and approaches the Greenwald density limit before disruption. And the row corresponding to the electronic density in heat map turns to be the brightest one. The result of single shot interpretation seems to be consistent with the physical expectation.

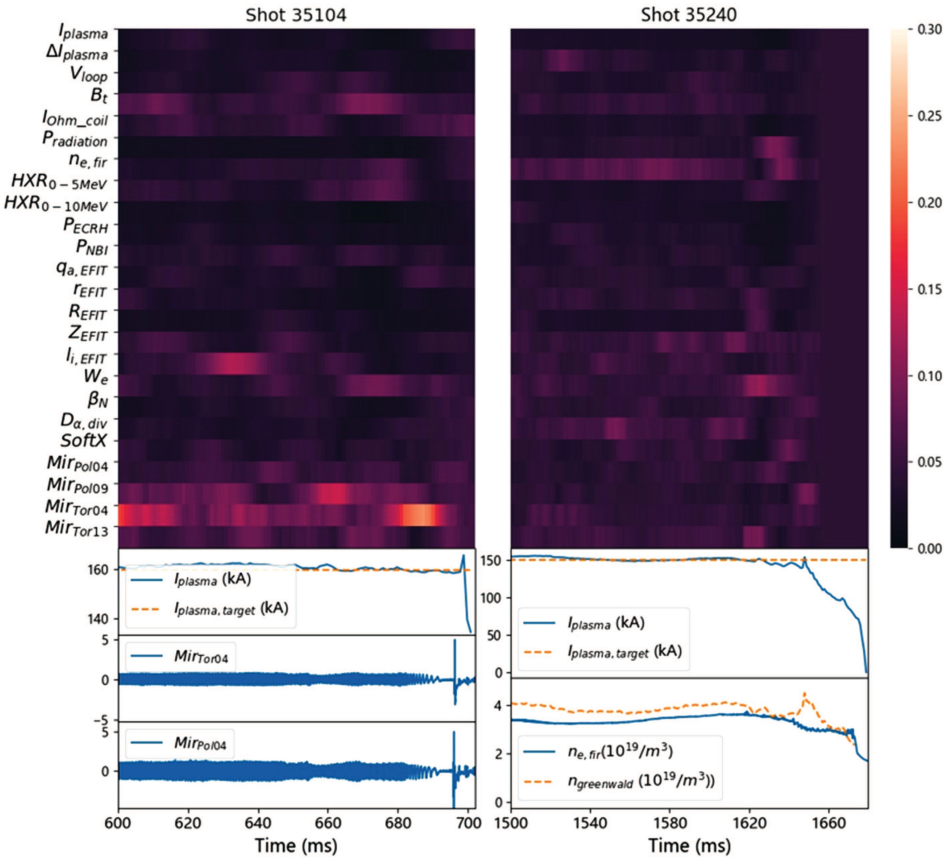


Fig. 6. Result of interpretation algorithm and related input signals of example shots. The left one is a lock-mode induced disruptive shot and the right one is a density limit induced disruptive shot. The figure is adapted from Fig. 3 in Ref. [17].

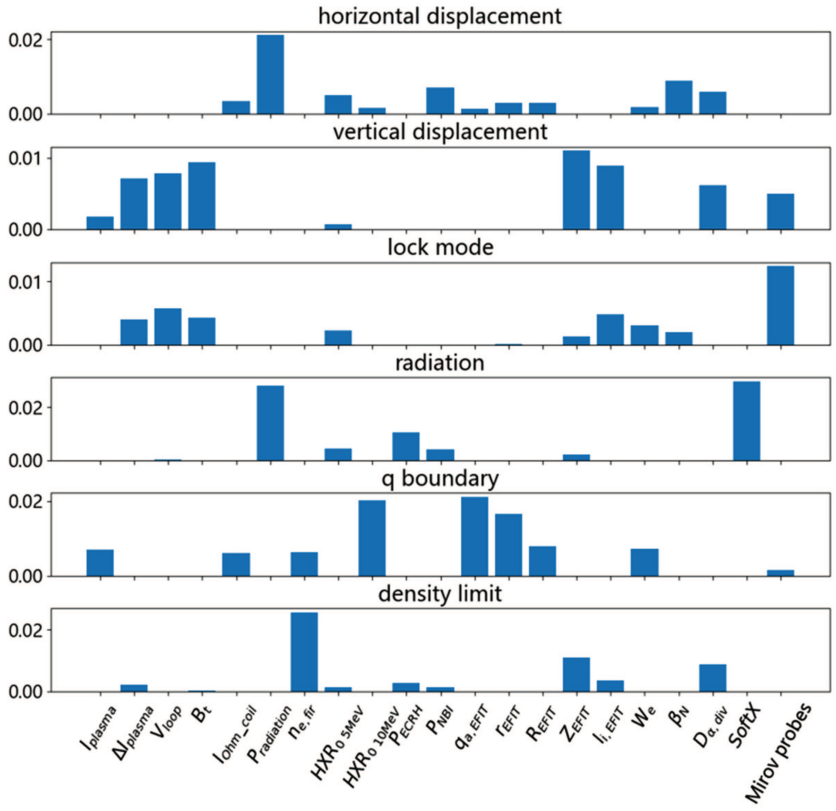


Fig. 7. Averaged correlation distribution of input signals on each disruption type in HL-2A. Each subplot corresponds to a type of disruption. The x-axis is the input signal list and the y-axis is the averaged correlation between the input signal and a certain type of disruptions.[17]

Statistical analysis isalso implemented to further validate the method. As a support, 613 shots in HL-2A are manually labeled into 6 disruption types, i.e., horizontal displacement, vertical displacement, radiation collapse, density limit, mode locking and low safety factor on plasma boundary. The averaged correlation distribution on each disruption type is calculated and shown in Fig. 7. As expected, the most important signals for vertical displacement, lock mode, radiation, low safety factor on boundary and density limit induced disruptions are Z_{EFIT}, Mirnov probes, $P_{\mathrm{radiation}}$/SoftX, $q_{\mathrm{a,EFIT}}$ and $n_{\mathrm{e,fir}}$, 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.

4.3. Disruption cause recognizer

An important application of interpretable disruption prediction is the active disruption avoidance. As a sanity check of this technique routine, a disruption cause recognizer is developed based on the previous interpretation method. When a trigger signal is sent by disruption prediction algorithm, the recognizer will take the correlation distribution within the last 30 ms as input and try to classify the disruption types listed in Subsection 4.2. Note that the data from type of low safety factor on plasma boundary are too few to support the training of model, so this type is abandoned here. The rest dataset contains 605 shots and 5 disruption types. A naive Bayes classifier is trained on this dataset. It can classify the 5 types of disruptions with an accuracy of 71.2% in 5-fold cross validation. Figure 8 gives the algorithm's confusion matrix during 10-fold cross validation. According to the result, it seems possible to develop a disruption cause recognizer and take measures to eliminate the disruption causes and avoid the disruptions in the future.

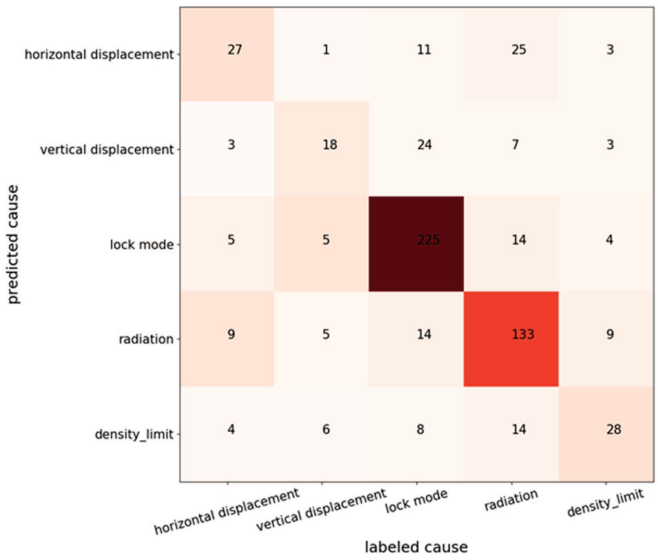


Fig. 8. Confusion matrix of the disruption cause recognizer. The x-axis represents the manually labeled disruption cause. And the y-axis refers to the predicted cause. So the numbers in the blocks on the diagonal line from top left to bottom right are the shot numbers that are correctly predicted, and vice versa.[17]

5. Few-shot learning on future devices

Since disruption prediction techniques are investigated to solve the problem of future fusion reactors, the challenge of few-shot learning becomes extremely important. Considering the unacceptable result of disruptions, future devices can only provide very limited amount of data to support the development of disruption prediction algorithm.[20] There are mainly two pathways to solve the problem within the deep learning paradigm. One is to develop a transferable algorithm, which learns data from existing tokamaks and can be adapted to future devices. Examples for this routine could be the adaptive learning algorithms,[21] FRNN with a glimpse of future device,[13] scenario adaptive learning[22] and open world learning.[23] In Subsection 5.1, a new method called device adversarial neural network (DANN) is tried and performs well when it is trained on HL-2A, J-TEXT and adapted to HL-2M. The other routine is to train a reliable algorithm with as fewer training data as possible, with the help of physical prior knowledge, some examples has been implemented in DIII-D[24] and J-TEXT.[25] Subsections 5.2 and 5.3 propose two possible physical inductive biases, i.e., plasma equilibrium equation and disruption related instabilities, that could be integrated into the disruption prediction algorithm. The future works in HL-2A will focus on these two inductive biases and preliminary works are implemented to validate the feasibility of these methods.

5.1. Device adversarial neural network

The researches in Refs. [13,14] prove that training with mixed dataset, which consists of a large amount of data from 'existing' devices and a few data from 'future' devices is a possible way to develop a cross-tokamak algorithm. DANN makes a further step on top of this method, which uses adversarial training strategy to extract transferable features on multiple devices on purpose. Figure 9 demonstrates the structure of DANN. The main difference between DANN and normal neural networks is that it has two output branches. One branch will predict the disruptions as normal, and the other branch will recognize the source device of the input data. Note that there is a gradient reversal layer in the second branch, which will opposite the gradients back propagated from the second branch during the training of neural network. Therefore the neural network will try to predict the disruptions as accurately as possible and try to recognize the source device of data as inaccurately as possible. In other words it will recognize the disruptions by common features among the multiple devices.

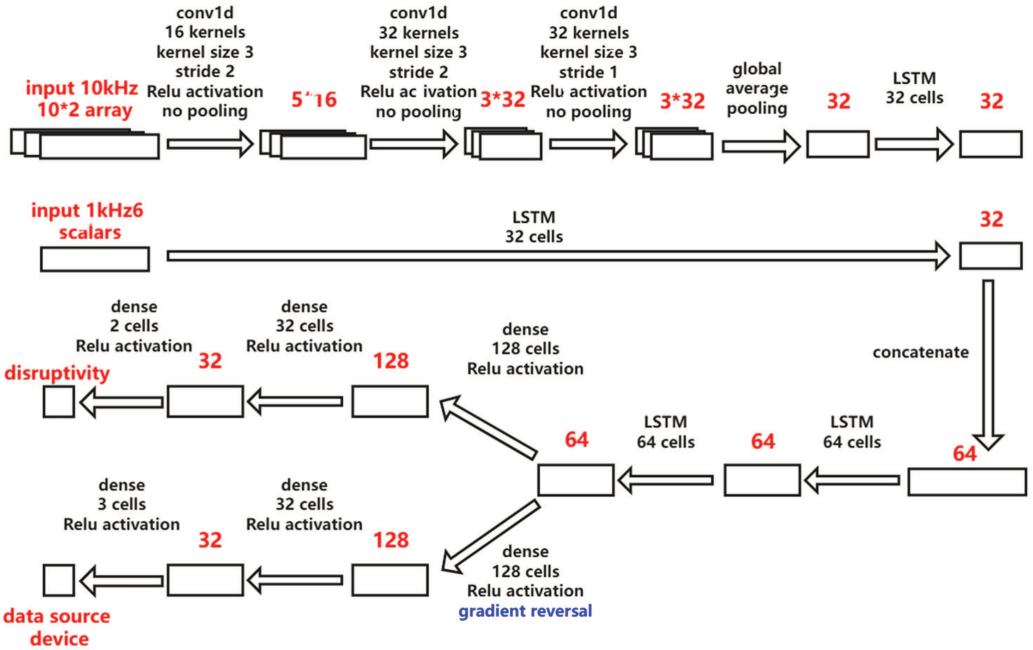


Fig. 9. Structure of the device adversarial neural network. The array size of input, output and middle nodes are marked with red texts. The neural network layer parameters are marked with black texts. The position of gradient reversal layer, which is the key of DANN, is marked with blue texts. The idea is based on the research in Ref. [26].

DANN is trained on HL-2A/J-TEXT/HL-2M mixed datasets. The main information of the mixed dataset is shown in Tables 4 and 5. The algorithm gets an accuracy of 88.9% with only 44 training shots from HL-2M.[19] However, the testing set provided by HL-2M is pretty limited now, 27 shots. More data are required to implement a reliable enough test on DANN.

Table 4. Shot numbers for training, validation, and testing of HL-2M's DANN disruption predictor. D means disruptive and ND means non-disruptive.[19]

Device	Training		Validation		Testing	
	D	ND	D	ND	D	ND
HL-2A	1532	4317	364	470	–	–
J-TEXT	1477	4091	415	1221	–	–
HL-2M	25	6	17	6	18	9

Table 5. The input signal list of HL-2M's DANN disruption predictor.[19]

Signal name	Sample rate (kHz)	Physical meanings	Signal tag		
			HL-2M	J-TEXT	HL-2A
Ip	1	plasma current	IP_2M	ip	IP
Target_Ip	1	target plasma current	ccIP_2M	oh_set	ccIP
Bt	1	toroidal magnetic field	BT1_2M	bt	Bt
Density	1	density of electrons	AMW_INT1_02	LIN_DEN_CH06	Density1
Dh	1	horizontal displacement	dh_2M	dx	FluxDh
Bolometer	1	power of bolometer radiation	DBOL_10	AXUV_CA_08	BOLU10
Mir_Probe_A	10	a pair of symmetric poloidal Mirnov probes	Mpol_17_24	MA_POL_CA01T	Mpol_04
Mir_Probe_B	10		Mpol_17_51	MA_POL_CA19T	Mpol_13

5.2. Parameter evolution equations

Another possible way to relieve the lacking of training data on future devices is to introduce more physical information. In deep learning paradigm, the physical information can be seen as the prior knowledge in the neural networks. And the research in this chapter aims to make plasma equilibrium equations into the neural network's prior knowledge. This purpose can be realized by two steps. Firstly, train a plasma parameter prediction algorithm, based on the parameter evolution equations. Secondly, embed the information hidden in parameter prediction algorithm into the disruption prediction algorithm.

The plasma parameter prediction algorithm predicts the evolution of electron temperature (Te), electron density (Ne) and horizontal displacement (Dh) of HL-2A according to the control target, control actuators and the plasma state. Table 6 shows part of the input and output list of the algorithm, three empirical equilibrium equations are hidden in the design of the signal list, such as the equation for Te given here. The input and output lists of Ne and Dh are designed with similar principles. By training the neural network to do the parameter prediction task, the equation can be embedded into the neural network, and could possibly be brought into the disruption prediction algorithm in the future.

Table 6. The hidden equation for electron temperature of the parameter prediction algorithm.

Input	Explaination	Hidden equation	Output
I_{plasma}	Basic plasma parameter	$\frac{\mathrm{d}T_e}{\mathrm{d}t} = k\frac{\mathrm{d}n_e}{\mathrm{d}t}T_{\text{puffing}} + \lambda(P_{\text{ohm}} + P_{\text{aux}} - P_{\text{rad}} - P_{\text{trans}})$	T_e (electron temperature measured by ECE, normalized to $[-1, 1]$)
P_{ECRH} P_{NBI} P_{LH} $I_{\text{Ohm_coil}}$	Ohmic and auxiliary heating powers, which should be positively related to the electron temperature		
$P_{\text{radiation}}$ $\text{Mir}_{\text{Pol13}}$	Two main loss mechanism of thermal energy, namely, radiation and transport. The radiation power is measured by bolometer array, while the transport process is kind of hard to measure. Therefore, Mirnov probe signal is used as a compromised alternative, which could be indirectly related to the transport level through the MHD physics.		
$\text{Gas}_{\text{puffing,FF}}$ $\text{Gas}_{\text{puffing,FB}}$ $\text{Gas}_{\text{SMBI,FB}}$	Gas puffing and supersonic beam injection brings cold density source into the plasma and thus is related to the equilibrium of electron temperature, too.		
R_{EFIT} Z_{EFIT} B_t $n_{e,\text{fir}}$	The election temperature is measured by Electron Cyclotron Emission (ECE) diagnostic, whose data is also influenced by plasma's position and shape, toroidal field and electron density.		

The algorithm is based on a encoder–decoder model. The encoder extracts the characteristics of the inputs and converts the inputs into the intermediate state matrix. The decoder reconstructs the intermediate matrix and restores the encoder outputs to the signals to be predicted. The algorithm is trained by 1500 shots and tested with 200 shots. The mean absolute errors (MAE) of Ne, Te and Dh on the testing set are 0.0393 (baseline: 0.1036), 0.0359 (baseline: 0.0561) and 0.0506 (baseline: 0.0738) respectively, while the baseline is evaluated by the difference between actual parameters and the delayed parameters in the inputs. The typical result is shown in Fig. 10, the output of algorithm is very close to the actual signals. It seems that the neural network model has correctly learned the plasma equilibrium equations.

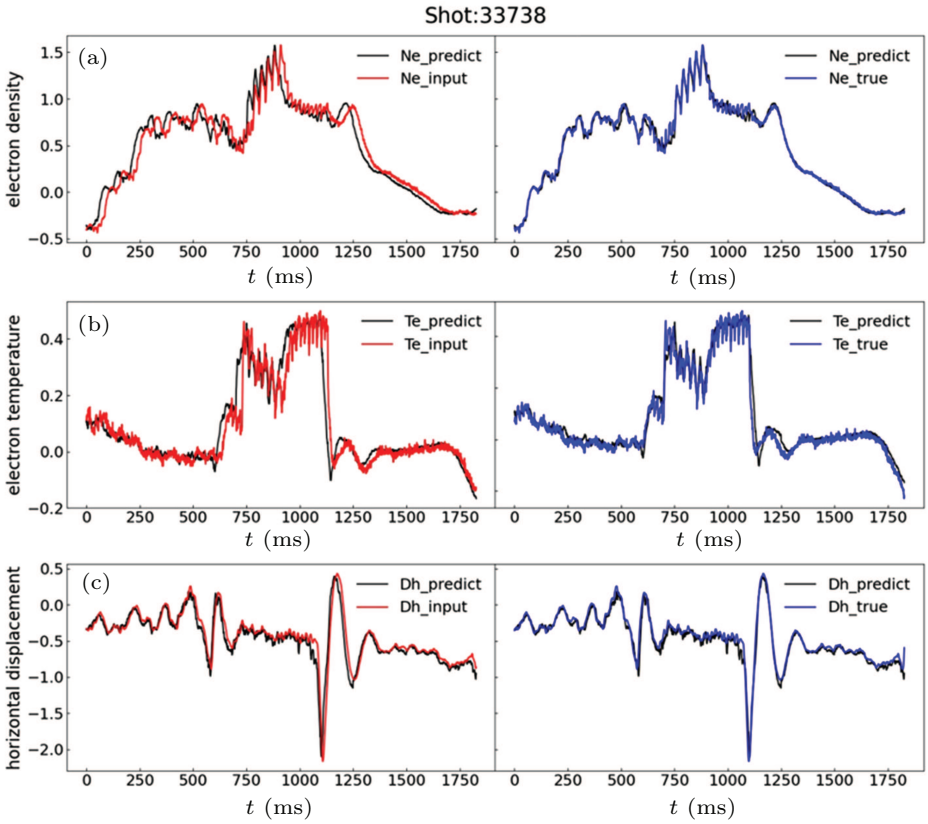


Fig. 10. Panels (a), (b) and (c) are electron density, electron temperature and horizontal displacement, respectively. On the left is the comparison of algorithm inputs and target outputs, where an obvious delay of ΔT can be observed. Here ΔT is 30 ms for Ne and Te, while it is 10 ms for Dh. On the right is the comparison of predicted parameters and target outputs, they are much closer than the situation in the left subfigures.

The second step, namely, to embed the prior knowledge in parameter prediction algorithm into the disruption prediction algorithm is still ongoing. A simple implementation is to concatenate the intermediate matrix given by encoder into the inputs of disruption prediction algorithm. With this auxiliary input, the AUC of disruption prediction algorithm could be promoted by 4%.

5.3. Disruption related instabilities

Plasma instabilities are also helpful prior knowledge for disruption prediction.[2] Therefore, researches on recognizing instabilities with neural network are also implemented in HL-2A. For example, an CNN+LSTM neural network is developed to recognize fishbone mode. The algorithm takes the signal of magnetic fluctuations measured by Mirnov coil during a 1 ms window with sample rate of 1 MHz, i.e., a 1000 length vector. The vector is firstly dealt with by mean-std normalization. Then a Fourier transformation with frame length of 1000 and frame step of 100 is implemented to get a 501×91 spectrum. Since the fishbone in HL-2A usually has a frequency between 5–35 kHz, only the spectrum in this frequency range is reserved. Finally the 30×91 spectrum is fed into a neural network as shown in Fig. 11. After training on 770 shots, the neural network can real-timely recognize if there is a fishbone instability in the plasma with an accuracy of 95.1%. Figure 12 presents the algorithm's output during an example shot. Similar works has been implemented on tearing modes, sawtooth, ELM and other instabilities. Future works will focus on combining these modules into the disruption prediction algorithm and check if this could alleviate the algorithm's demand for training data.

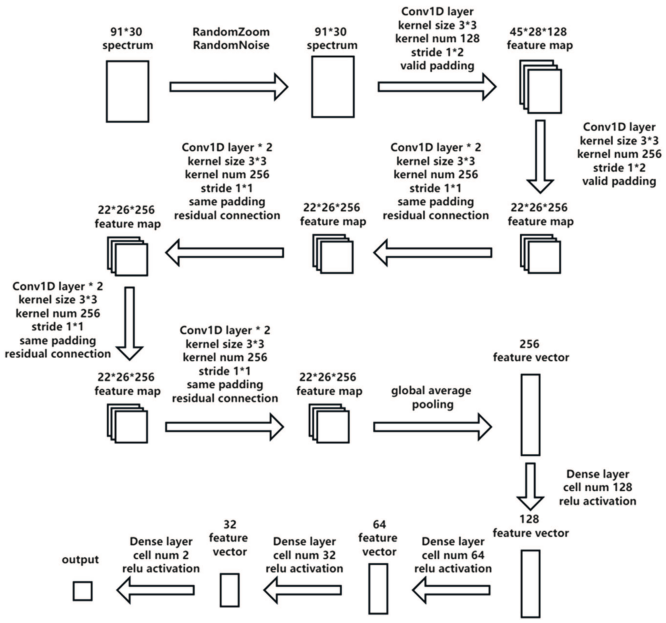


Fig. 11. Structure of the fishbone recognition neural network.

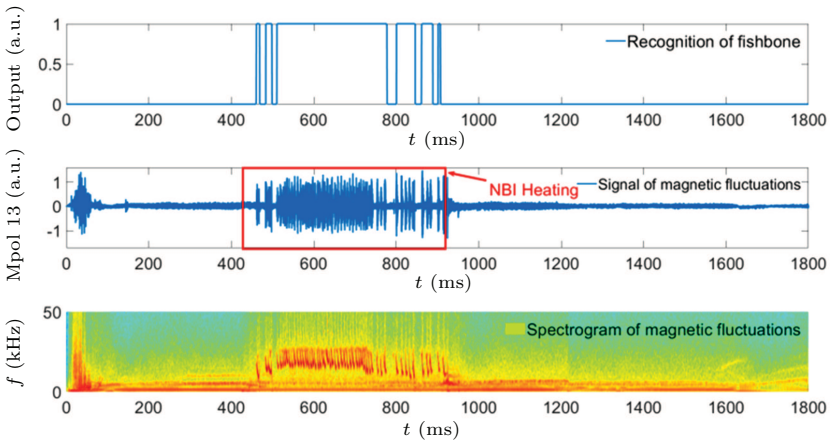


Fig. 12. An example shot of fishbone instability recognition algorithm.

6. Summary and future works

The deep learning-based disruption prediction algorithm has proved its powerful ability in HL-2A tokamak. With dedicated neural network design, optimized training strategy and novel interpretation method, the algorithm can realize high accuracy, good interpretability and priliminary cross-tokamak adaptation at the same time. Real-time disruption prediction and mitigation experiment is taken in HL-2A to validate the algorithm and the result proves its reliability. In general, deep learning can be considered as a sophisticated solution for disruption prediction in 'existing' tokamaks, both from a research and engineering perspective. Therefore the future works will go beyond the 'prediction on existing tokamaks'. There are mainly 2 pathways, i.e., 'prediction on future tokamaks' and 'prevention on existing tokamaks'.

To predict the disruption in 'future' tokamaks, the preliminary research in Section 5 will be carried on. Domain adaptation techniques and physical inductive bias will be introduced to alleviate the algorithm's demand for training data. HL-2M, a newly built tokamak which has realized the first plasma in 2020, has promoted its plasma current to 1 MA recently and will provide a great testing platform for these researches.

The other pathway is to prevent the disruptions based on interpretable disruption prediction algorithm. A correlation analysis has been realized on HL-2A's algorithm in Section 4. However, the method needs to go beyond the correlation and realize causality analysis to find the root cause from a bunch of correlated phenomena. A cause-and-effect diagram needs to be designed with consideration of the callable control actuators. Therefore, the root cause can be matched to the control measures to support the disruption prevention.

In general, deep learning provides a potential solution for the disruption prediction, prevention and mitigation on future tokamaks. The basic feasibility has been proved in HL-2A and many other tokamaks in recent years. Some bottleneck problems, such as cross tokamak adaptation, disruption cause recognition still exist, but are being investigated and preliminary solved. It can be expected that there will be more breakthroughs in this direction in the next few years.

Acknowledgements

Project supported by the National MCF R&D Program of China (Grant Nos. 2018YFE0302100 and 2019YFE03010003). The authors wish to thank all the members at South Western Institute of Physics for providing data, technique assistance and co-operating during the experiment.

References

[1] Wesson J 2011 *Tokamaks*, fourth edition (Oxford University Press) pp. 15–27
[2] ITER Physics Expert Group on Disruptions 1999 *Nucl. Fusion* **39** 2251
[3] Taylor P L, Kellman A G, Evans T E, *et al* 1999 *Phys. Plasmas* **6** 1872
[4] Nardon E, Hu D, Hoelzl M, *et al* 2020 *Nucl. Fusion* **60** 126040
[5] Vries P C, Johnson M F, Alper B, *et al.* 2011 *Nucl. Fusion* **51** 053018
[6] Vries P C, Johnson M F, Alper B, *et al.* 2011 *Nucl. Fusion* **51** 053018
[7] Rea C, Montes K J, Erickson K G, *et al.* 2019 *Nucl. Fusion* **59** 096016
[8] Ratta G A, Vega J, Murari A, *et al.* 2010 *Nucl. Fusion* **50** 025005
[9] Aymerich E, Sias G, Pisano F, *et al.* 2022 *Nucl. Fusion* **62** 066005
[10] Zhong Y, Zheng, Chen Z Y, *et al.* 2021 *Plasma Phys. Control. Fusion* **63** 075008
[11] Zheng W, Wu Q Q, Zhang M, *et al.* 2020 *Plasma Phys. Control. Fusion* **62** 045012
[12] Guo B H, Chen D L, Shen B, *et al.* 2021 *Plasma Phys. Control. Fusion* **63** 115007
[13] Harbeck J K, Svyatkovskiy A, Tang W, *et al.* 2019 *Nature* **568** 526
[14] Zhu J X, Rea C, Montes K J, *et al.* 2021 *Nucl. Fusion* **61** 026007
[15] Ma R and Wang A K 2010 *Nucl. Fusion Plasma Phys.* **30** 37
[16] Yang Z Y, Xia F, Song X M, *et al.* 2020 *Nucl. Fusion* **60** 016017
[17] Yang Z Y, Xia F, Song X M, *et al.* 2021 *Nucl. Fusion* **61** 126042
[18] Yang Z Y, Xia F, Song X M, *et al.* 2022 *Fusion Eng. Des.* **182** 113223
[19] Yang Z Y, Xia F, Song X M, *et al.* 2022 *Proceeding on 48th EPS conference on plasma physics*, June 27th–July 1st, 2022, Amsterdam, Holland, p4a.101
[20] Vries P C, Pautasso G, Humphreys D A, *et al.* 2016 *Fusion Sci. Tech.* **69** 471
[21] Murari A, Rossi R, Peluso E, *et al.* 2020 *Nucl. Fusion* **60** 056003
[22] Zhu J X, Rea C, Granetz R S, *et al.* 2021 *Nucl. Fusion* **61** 114005
[23] Gelfusa M, Murari A, Vega J, *et al.* 2021 *4th IAEA Technical Meeting on Fusion Data Processing, Validation and Analysis*, November 29th–December 6th 2021, Chengdu, China, ID: 4
[24] Shen C S, Zheng W, Ding Y H, *et al.* 2023 *Nucl. Fusion* **63** 046024
[25] Zhu J X, Rea C, Granetz R S, *et al.* 2023 *Nucl. Fusion* **63** 046009
[26] Ganin Y, Ustinova E, Ajakan H, *et al.* 2016 *Journal of Machine Learning Research* **17** 1