IMPLEMENTATION OF AI/DEEP LEARNING DISRUPTION PREDICTOR INTO A PLASMA CONTROL SYSTEM

William Tang, Princeton University, PPPL
Princeton Plasma Physics Laboratory, Princeton, New Jersey, 08543 USA
Email: wtang@pppl.gov

Ge Dong[1], Jayson Barr,[2] Keith Erickson[1], Rory Conlin[1], Dan Boyer[1], Julian Kates-Harbeck[1], Kyle Felker[1], Alexey Svyatkovskiy[1], Eliot Feibush[1], Joseph Abbatte[1], Mitchell Clement[1], Brian Grierson[1], Raffi Nazikian[1], Nik Logan[1], Zhihong Lin[3], David Eldon[2], Auna Mosr[2], Cristina Rea[4], Mikhail Maslov[5]

1) Princeton Plasma Physics Laboratory, Princeton, New Jersey, 08543 USA; 2) General Atomics, San Diego, California 92186, USA; 3) University of California Irvine, Irvine, California 92717, USA; 4); 4 MIT, PSFC, 77 Massachusetts Ave., Cambridge, MA 02139; 5) EUROfusion Consortium, JET, Culham Science Centre, Abingdon, OX14 3DB, UK

Abstract

State-of-the-art deep-learning disruption prediction models based on the Fusion Recurrent Neural Network (FRNN) introduced in the NATURE (2019) publication have been further improved. The paper reports on new capability of the AI/DL software to output not only the "disruption score," as an indicator of the probability of an imminent disruption, but also a "sensitivity score" in real-time to indicate the underlying reasons for the imminent disruption. This provides valuable physics-interpretability for the deep-learning model results and associated guidance for control actuators when implemented into a modern Plasma Control System (PCS). The advance is a significant step forward in moving from modern DL disruption prediction to real-time control and brings novel AI-enabled capabilities relevant for application to the future burning plasma ITER system. Results presented used large amounts of data from JET and DIII-D vetted in the earlier NATURE publication. In addition to "when" a shot predicted to disrupt, this paper addresses reasons "why" by carrying out sensitivity studies. FRNN is accordingly extended to use many more channels of information, including measured DIII-D signals such as (i) "n1rms" that is correlated with the n =1 modes with finite frequency, including neoclassical tearing mode and sawtooth dynamics; (ii) the bolometer data indicative of plasma impurity content; and (iii) "q-min" – the minimum value of the safety factor relevant to the key physics of kink modes. When integrated into the deep learning FRNN software, clearer identification of physics responsible for the disruption events was enabled with associated relevant guidance for control actuators. In providing a "disruption score" together with a "sensitivity score" for each physics-connected channel, the present investigations of disruption subcategories for relevant physics channels can provide more precise and direct information for the control actuators in a plasma control system.

1. INTRODUCTION

State-of-the-art deep-learning disruption prediction models based on the Fusion Recurrent Neural Network (FRNN) [1] have been further improved. Here we report the new capability of the software to output not only the "disruption score," as an indicator of the probability of an imminent disruption, but also a "sensitivity score" in real-time to indicate the underlying reasons for the imminent disruption. As an indicator of possible causes for future disruptions, the "sensitivity score" can provide valuable physics-based interpretability for the deep-learning model results, and more importantly, provide targeted guidance for the control actuators when implemented into any modern Plasma Control System (PCS). This achievement represents a significant step forward since the 2018 IAEA meeting in moving from modern deep-learning disruption prediction to real-time control that brings novel AI-enabled capabilities needed for the future burning plasma ITER system.

The main findings in this paper help address the basic issue/perception that advanced Machine Learning/Deep learning methods are generally hard to interpret. Results presented here are of course supportable by actual data from JET and DIII-D with much of such data having been previously published/vetted in the prominent deep learning NATURE paper [1]. Moving beyond this work on tokamak disruptions, the current paper addresses and answers in addition to "when" a shot is going to disrupt, some compelling reasons to explain "why" it disrupts by carrying out sensitivity studies

A new scheme is introduced in which real-time control of actuators can be advanced by AI-enabled disruption predictors. Since these deep learning capabilities were developed by using modern programming languages (i.e., Python) to implement the "Keras" algorithmic scheme (explained in [1]), it has become additionally necessary to develop a "Keras2c" converter to enable integration of the AI-based predictor into a real-time plasma control system

(e.g., for DIII-D) which is written in the much older C-language. Associated details will be further explained later in Section 2.1 of this paper. It is important to keep in mind that the cited NATURE paper represents the first adaptable predictive DL software trained on leadership class supercomputing systems to deliver accurate predictions for disruptions across different tokamak devices (DIII-D in the US and JET in the UK). It features the unique statistical capability to carry out efficient "transfer learning" via training on a large database from one experiment (i.e., DIII-D) and be able to accurately predict disruption onset on an unseen device (i.e., JET). In more recent advances, the FRNN inference engine has been deployed in a real-time plasma control system on the DIII-D tokamak facility in San Diego, CA. This opens up exciting avenues for moving from passive disruption prediction to active real-time control with subsequent optimization for reactor scenarios.

The workflow for the FRNN software can be readily extended to explore the use of many more channels of information. For example, DIII-D signals that are known to be highly relevant physics-based channels are: (i) "n1rms" – a signal correlated with the activities of the n =1 modes with finite frequency, including the neoclassical tearing modes (NTMs) and sawtooth; (ii) the bolometer data that reflects the impurity content of the plasma; and (iii) "q-min" – the minimum value of the safety factor which directly relates to important physics such as the kink modes. These considerations have motivated including the associated channels directly into the deep learning workflow with the goal of clearer identification of the physics most responsible for the dangerous disruption events with associated guidance for the control actuators. The potential for significant improvement over existing traditional algorithms targeting these signals for plasma condition and disruption control comes from the fact that our AI/deep-learning models are set up for carrying out supercomputing-enabled hyperparemter tuning enhancements of satistical accuracy for complex physical systems with huge feature size without the necessity of "feature engineering." This enables the capability to deliver predictions for unseen conditions, such as new plasma parameters associated with projected larger devices. Moreover, as an indicator of possible causes for future disruptions, the distribution of the "sensitivity score" can provide valuable physics-based interpretability for the deep-learning model results, and more importantly, provide targeted guidance for the control actuators when implemented into any modern PCS. Progress toward this goal represents a significant step forward in moving from modern deep-learning disruption prediction to real-time control that brings novel AI-enabled capabilities with significant beneficial features for deployment in the future on the burning plasma ITER system. Results indicate, for example, that the core radiation power and the familiar MHD safety factor at the radial location near the plasma periphery (*q-95*) can represent sensitive channels responsible for associated disruption prediction for specific cases of interest.

This paper emphasizes that the FRNN software can be readily extended to using many more channels of information. In particular, there are DIII-D signals that are known to be highly relevant physics-based channels, which include: (i) "n1rms" – a signal correlated with the activities of the n =1 modes with finite frequency, including the neoclassical tearing modes (NTM's) and MHD sawtooth dynamics; (ii) the bolometer data that reflects the impurity content of the plasma; and (iii) "q_{min}" – the minimum value of the safety factor which directly relates to important physics such as the kink modes. These considerations have strongly motivated the inclusion of associated channels directly into the improved performance development of the deep learning FRNN software with the goal of clearer identification of the physics most likely responsible for the dangerous disruption events with associated guidance for the control actuators. The potential for significant improvement over existing traditional algorithms targeting these signals for plasma condition and disruption control comes from the fact that AI/deep-learning models have the distinct advantage of being able to greatly enhance the predictive accuracy via modern hyperparameter tuning with associated training carried out on path-to-exascale supercomputers at HPC facilities such as ORNL, ANL, and LBNL in the US. This enables the capability to deliver predictions for as-yet-unseen conditions that can arise, such as new plasma parameters including new DIII-D experiments and those associated with future larger devices such as ITER.

Overall, the key point made in this paper is that when more physics related channels are included, key insights can be gained on the mechanisms contributing to disruptions. Accordingly, in addition to providing a "disruption score," the present studies compute a "sensitivity score" for each physics-connected channel (as illustrated in Fig. 2). In addition to studying the physics in subcategories of disruptions, these "sensitivity scores" for each channel can in turn provide guidance to the PCS with more precise and direct information for the control actuators. Moreover, another key advantage of deep-learning enabled predictive capabilities is the ability to carry out forecasts significantly earlier in the evolution of the plasma state under consideration. For example, in exploratory studies carried out here with the "n1rms" signal information included, the alarm time for FRNN disruption prediction can be estimated around 100 ms as compared to the 30 ms estimates noted in the cited Nature paper.

As background for the present studies, we note that in toroidal plasma devices, disruptions are large-scale plasma instabilities that release the plasma stored energy and diminish the plasma current within a very short time-scale [7]. The large energy and particle flux involved can seriously damage the experimental devices, especially when stored energy increases in high-performance plasma experiments (shots) in modern tokamaks such as DIII-D [8] and JET [9], and future tokamak devices such as ITER [10]. In addressing this long-standing challenge, neural networks have been considered for decades [11]. More recently, deep learning models based on the Long-Short Term Memory (LSTM) recurrent neural network (RNN) have achieved breakthrough results [1], for cross-machine predictions with the aid of modern High-Performance-Computing (HPC) capabilities [12-13].

In this paper, we present the first results of the real-time implementation of a FRNN LSTM-based deep-learning model into the DIII-D PCS. The real-time computation performance of the FRNN inference engine during DIII-D start-up runs is proven to be compatible with the PCS requirements, which demonstrates that FRNN deep-learning models are entirely capable of real-time disruption prediction tasks to aid disruption control. We also highlight here some recent off-line FRNN results, including a new training scheme with more physics-based signals to improve FRNN disruption prediction capabilities and a new FRNN software suite to compute real-time "sensitivity-scores" for the interpretation and of FRNN deep-learning model disruption predictions. The remainder of the paper is organized as follows: in section 2, we provide details of the implementation of the FRNN deep-learning based model into the DIII-D PCS; in section 3, we discuss new FRNN training and disruption prediction results when new physics related signals are included as inputs; in section 4, we introduce the design and output of the "sensitivity scores", and in section 5, we summarize the recent advances and planned future developments of the FRNN software suite.

2.0 IMPLEMENTATION OF FRNN DEEP-LEARNING-BASED MODEL INTO DIII-D PCS

The DIII-D tokamak uses a general real-time plasma control system framework (PCS) created at General Atomics (GA) and shared with multiple other facilities, including the international long-pulse tokamaks KSTAR in Korea and EAST in China – as well as the spherical torus experiments NSTX-U in the US and MAST-U in the UK. This framework enables running feedback control algorithms on microsecond timescales in a highly deterministic fashion, complete with documented information on configuration, archival data, introspection, and other important capabilities [14]. The AI/deep learning FRNN software has now been integrated into this framework as a new category of algorithms. It has demonstrated successful operation encompassing a significant number of DIII-D shots in the past year. This implementation consists of four parts: (i) pre-shot configuration; (ii) real-time data collection; (iii) processing through a Keras2c interpreter (see Section 2.1 below); and (iv) collection of results with associated documentation/publication. The PCS includes a complete user interface allowing an operator to easily choose configuration parameters specific to an algorithm. In the case of FRNN, the required configuration is a list of normalizing factors to apply to each input. These factors are set before the shot and applied during the real-time data collection phase. In particular, this phase amalgamates heterogenous data from multiple sources during each real-time cycle that includes diagnostics, sensor measurements, and internal calculations from other algorithms. Those values can then be adjusted/manipulated in various ways, including the application of pre-shot normalizing factors to match the offline functions used when training the model. It is important to note here that *functional parity is critical to ensure that the values submitted during real-time correlate to values trained offline*. Finally, the collection of data inputs are inserted into a predefined Keras2c input data format for use in the generic Keras processor shared by multiple algorithms. The Keras processor produces a predictive result which is then stored post-shot and published in real-time for any interested consumer. More specifically, "Keras2c" is a python/C library for converting complex Keras/Tensorflow neural networks such as the AI/deep learning FRNN software to real-time compatible C code. A python script parses the trained neural network to extract the necessary parameters and determine the connectivity between layers and nodes. It generates a custom C function to duplicate the forward pass through the neural network. The generated C code makes use of a small C-backend that re-implements the core functionality of Keras/Tensorflow in a real-time safe manner that allows ease of deployment into complex control systems such as the DIII-D PCS. "Keras2c" also automatically tests and verifies the correctness of the generated code. *The conversion and testing process of "Keras2c" is fully automated, providing a significant advantage over previous attempts to use neural networks within demanding control applications*. This advance enables avoiding the need to either use large non-deterministic software libraries or to code the entire network by hand, and thereby avoiding generating code that proves to be difficult to verify and maintain [3].

2.1 Current progress in real-time performance of FRNN

The AI/DL FRNN software has successfully initiated real-time runs on the DIII-D PCS, including establishing its own dedicated category to avoid potential conflicts with other algorithms. It has now demonstrated reproducible timing for representative real-time shots on the order of 1ms. Moreover, the "Keras2c" infrastructure has now been further upgraded to enable sharing the implementation across multiple algorithms (i.e., as many as four) to provide flexibility to address/remediate a number of integration issues that can arise within the PCS. Finally, it is significant to note that with respect keeping pace with attractive emerging technological advances, this AI/DL project has recently leveraged engagement with industry (NVIDIA and Concurrent) to build a unique real-time system using a new NVIDIA A100 GPU that has now been integrated into the DIII-D PCS. This can be expected in the near future to enable quicker and more efficient examination of potential benefits of deploying our current as well as possible new AI/DL algorithms.

3.0 PHYSICS-BASED SIGNALS FOR IMPROVING DISRUPTION PREDICTION

As mentioned in the previous section, we studied the effects of including more physics-based signals as inputs to our deep learning based models, and found that they can improve predictive capabilities. For example, the finite frequency n=1 mode amplitude is an important physical quantity, where n is the toroidal mode number. We note here that although "n1rms"— as well as "n2rms" and "n3rms"— are post-processed non-causal data, we use "n1rms" to demonstrate the importance of including these instability-related signals in deep-learning based models by *shifting the "n1rms" signal input in time by 20 ms to prevent the model from seeing any future information about the plasma.* In the future, our FRNN database can easily be extended to include additional physics information of interest contained in real-time signals → for example the outputs from rtNewSpec algorithms [5]. Nevertheless, "n1rms" can be useful for indicating the stability properties of various important plasma instabilities, including the kink-like modes and neoclassical tearing modes – that can eventually "lock" to the inner wall of the device and lead to disruptions.

When the n=1 finite frequency mode amplitude is included as an input channel in FRNN, the disruption prediction results can be improved significantly at both the low false positive rate regime and the high false positive regime, as shown in Fig. 1. We performed hyperparameter tuning, as introduced in [1], for FRNN models that are trained with and without the "n1rms" signal and reported the performance of the models that achieved the highest area under the ROC curve on the validation set respectively.

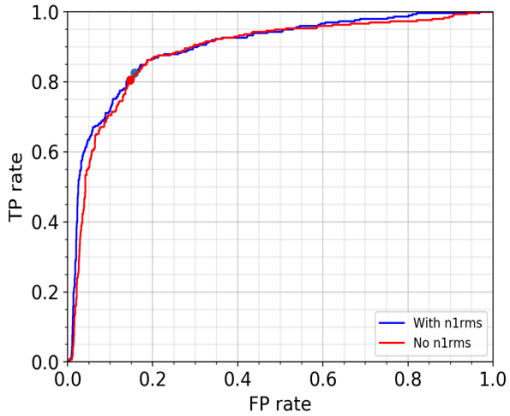

FIG. 1. Comparison of the ROC curve with and without the n=1 finite frequency mode amplitude ("n1rms")

More importantly, at the optimal alarm thresholds as indicated by the solid dots in Figure 3, the model trained with n=1 finite frequency mode signal can raise earlier disruption alarms than the model trained without the n=1 finite frequency signal. Both mean and median of the alarm time are increased by more than 100 ms. To demonstrate that the neural network can effectively learn information from the n=1 finite frequency mode signal and thus provide earlier disruption alarm, an example shot from DIII-D (shot #161362) is shown in the upper panel in Figure 4. At around 1.9s, the FRNN model with "n1rms" as an input channel raised the disruption alarm following the onset of the n=1 mode. Before 2s, the "n1rms" signal diminishes while the locked mode amplitude rises up. During this

time, the FRNN model trained with the "n1rms" signal provides continuous outputs of disruption alarms. The FRNN model trained without the n=1 mode signal raises a disruption alarm here around 40 ms before the actual disruption, around 200 ms later than the model trained with the n=1 mode signal.

With the other basic plasma quantities as input, the n=1 finite frequency signal usually does not confuse the model when the mode is not leading to disruptions. For example, in shot #170239, as shown in the lower panel of Fig. 2, although the neoclassical tearing mode appears at 2-3s, the FRNN output remains at a constant low level.

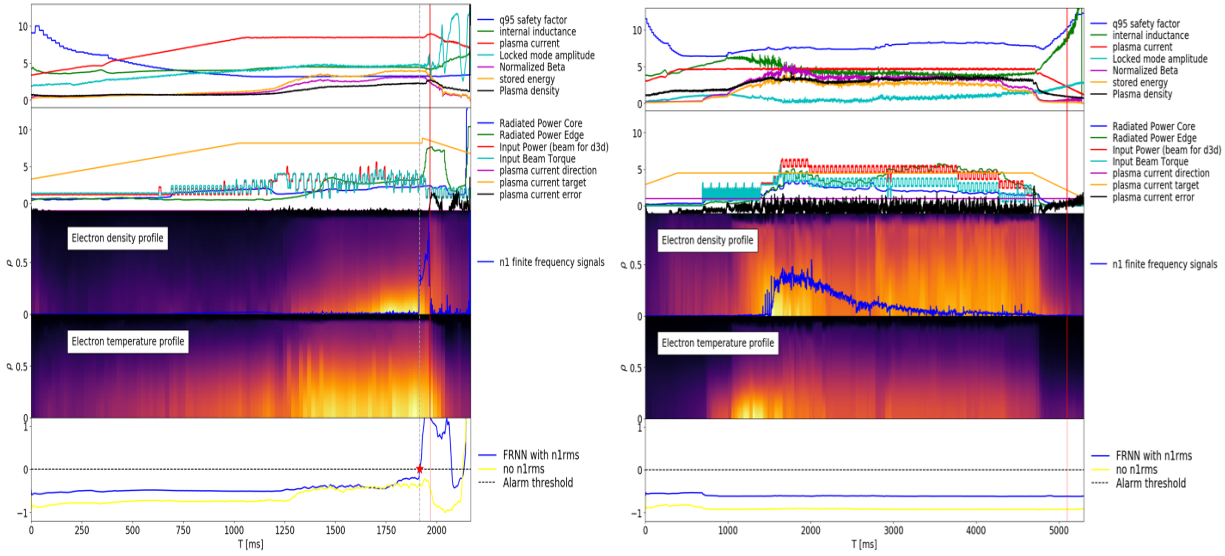

FIG. 2. DIII-D shot number 161362 in the left panel and DIII- shot number 170239 in the right panel. In each panel, the upper 4 subpanels show measured signals as FRNN input, and the bottom subpanel show FRNN model outputs

4.0 REAL-TIME SENSITIVITY STUDIES

To interpret the disruption predictive capabilities of the deep-learning-based model, we have developed sensitivity study schemes for individual test shots. These schemes can be implemented in real-time along with the regular FRNN model inference engine as introduced in Sec. 2. For each shot, the sensitivity study helps answer the question of 'why the neural network outputs a high disruption score and raises a disruption alarm at a given time?' More importantly, the results from the sensitivity study scheme can provide detailed indications of which physical quantities can provide relevant proximity guidance for disruptive scenarios, and this information may directly aid real-time control efforts.

4.1 Calculation of the sensitivity score

In Ref. [1], the authors provided results from studies examining the importance of a signal to illustrate the contribution of each such physical signal to the test results of the entire test database. In these signal importance studies, the model is re-trained for a "c" number of times with each physical signal excluded from the training and test database, where "c" is the number of physical signals, and the test result is reported in comparison with the baseline where all signals are included. In the present paper, we have included all signals during training. In the course of testing for each shot, we have performed inference "c" times in parallel – such that at each time, one physical signal is suppressed to output a "c" number of disruption scores. The sensitivity score of each signal is defined as the absolute difference between the baseline output (where full input is used) and the output where the signal of interest is suppressed in the test data. For the standard trained models, we suppress a signal by replacing that signal with a fixed typical value from non-disruptive shots. For noise-aware models which are trained with dropped out signals [5], and are 'familiar' with all-zeroes-signals, we suppress a signal by directly replacing with zeros. The outputs of these two schemes are generally qualitatively consistent, showing the robustness of the sensitivity study results with respect to the replacement values.

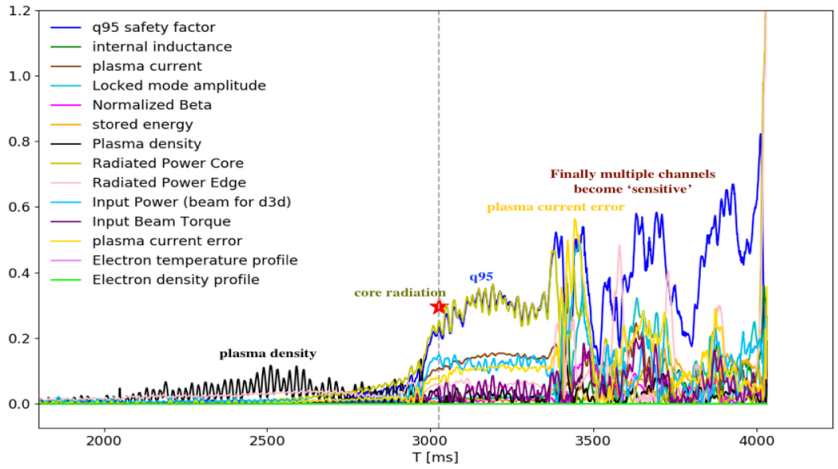

FIG 3. Evolution of the sensitivity score of the shot #164582

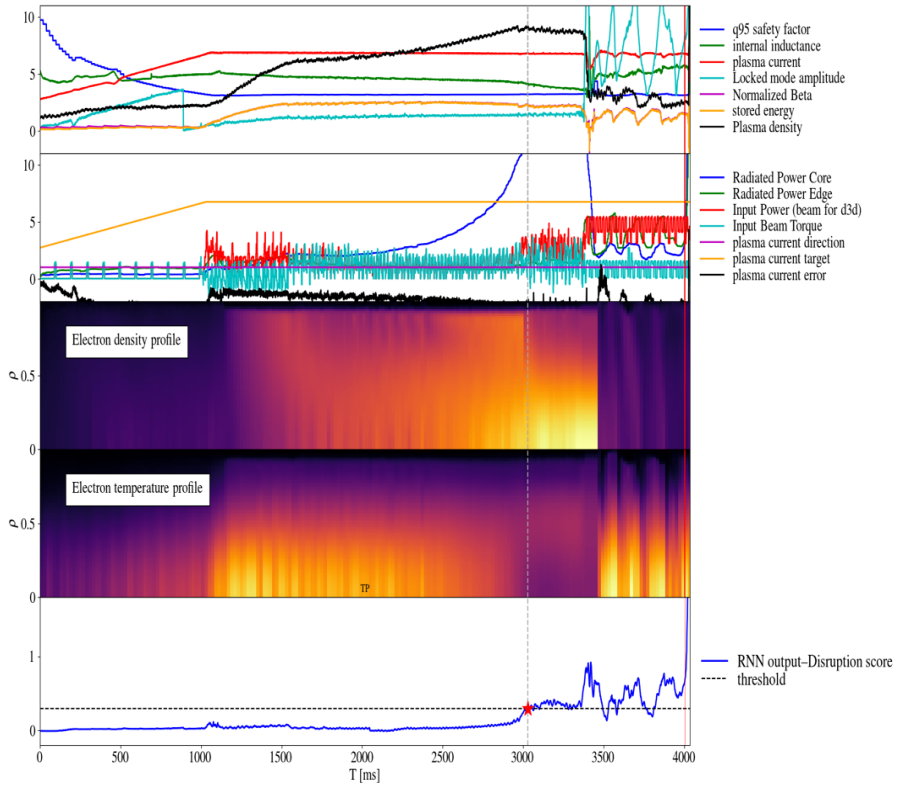

FIG 4. Evolution of each normalized physical signals for DIII-D shot #164582 in the upper 4 panels. The bottom panel shows the time history of the FRNN output

Fig. 3 shows an example of the sensitivity study result of DIII-D shot #164582, where the evolutions of the sensitivity scores of each signal are plotted as a function of time. At around 2 to 2.7 seconds, the disruption score slowly rise up, as shown in the last panel in Figure 6, and the plasma density is shown as the most sensitive channel. Experimentally during this time, the plasma density gradually rises due to impurity influxes, as shown in the first panel as a black line in Fig. 4. The influx of impurities is followed by the rise of the core radiation, which lead to disruption alarm at around 3s. At the disruption alarm time, the channels with the highest sensitivity scores are core radiation and q95. At round 3.5s, a large tearing mode starts to develop, possibly due to the high impurity level.

This in turn leads to mode locking as shown by the onset of the locked mode amplitude in the first panel of Fig. 3. Around this time, every channel becomes sensitive, and the sensitivity scores begin to change rapidly as the plasma current error rise up.

4.2 Sensitivity score from "Zero-value Replacement Procedure" for "noise-aware" models

In Fig. 5 we show the sensitivity study result for DIII-D shot #162975, using the "zero-value replacement procedure" for individual channels during inference studies using noise-aware models. It is illustrated that after 1.5 seconds, the disruption score begins to increase and raised a disruption alarm. The sensitivity score of different signals at the alarm time is shown in the lower panel of this figure as an indication of their contributions to the disruption alarm. The internal conductivity, q95, and plasma density are all sensitive channels that can indicate a general deterioration of plasma shape and plasma control. This interpretation is qualitatively consistent with the observation of experimental characteristics. At around 1.3 s, the plasma shape begins to become altered as the X point moves off-target. At 1. 45s, the tearing mode begins to develop -- which plausibly evolves into the locked mode at round 1.6-1.7s when the disruption alarm rises. In this shot, the locked mode amplitude is contaminated by stron5g noise from the extra "i-coil waveform."

It is important and significant to highlight the fact that in both of the shots that we have analyzed as examples here that although tearing modes which finally locked are important causes for plasma disruptions, the actual locked mode amplitude is not a sensitive signal for both of these shots. Since we did not use the "n1rms" signals as input for these studies, our results indicate that the deep neural network can process the basic plasma information effectively and that the interpretation of the outputs in the form of a real-time sensitivity study can provide early diagnostic information with associated guidance for plasma control and detailed disruption proximity analysis. Accordingly, such sensitivity studies are indeed potentially capable of contributing significantly to real-time disruption mitigation and avoidance investigations.

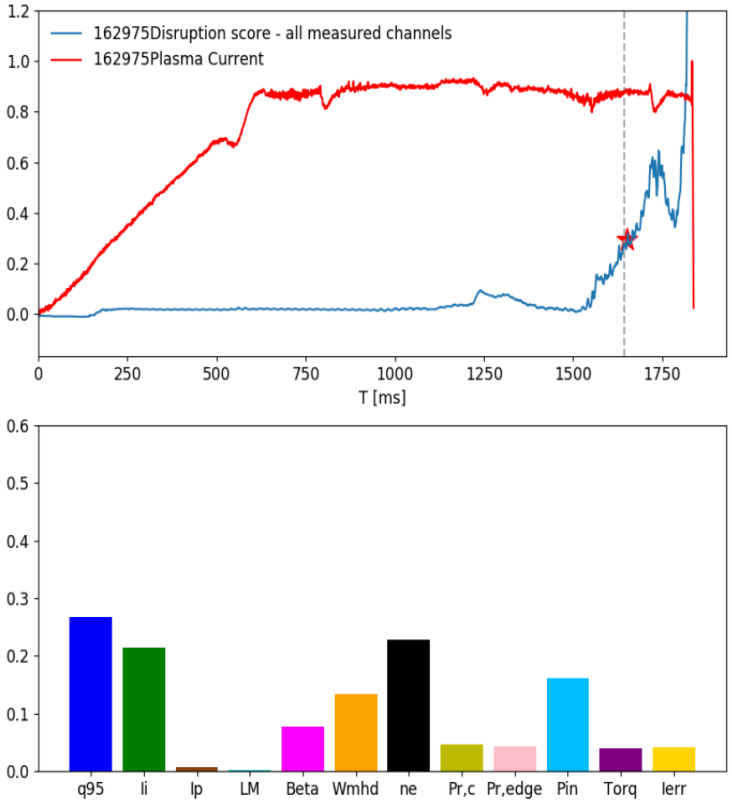

FIG. 5. DIII-D shot #162975. The upper panel shows the evolution of the plasma current (red line) and the FRNN output of the disruption score (blue line)

5.0 SUMMARY & ASSOCIATED FUTURE INVESTIGATIONS

In the present paper we have explained and provided details of the implementation of our AI/deep learning based models into the DIII-D plasma control system (PCS). This has included the introduction of a new method of interpreting results from deep neural networks using a sensitivity study methodology, The associated significance and implications for future plasma control systems have been carefully articulated. In ongoing and future investigations, we plan to extend and interconnect both aspects of this work by connecting the deep learning based model with inference engine implemented to an advanced workflow that integrates real-time sensitivity studies output to the proximity control architecture designed for handling major disruption causes in the DIII-D plasma control system [6]

REFERENCES:

[1] J. Kates-Harbeck, A. Svyatkovskiy, W.Tang, Nature 568 (7753), 526, 2019.
[2] P.C. de Vries, et al, Nucl. Fusion, 51, 053018, 2011.
[3] Conlin, Rory, et al. "Keras2c: A library for converting Keras neural networks to real-time compatible C." *Engineering Applications of Artificial Intelligence* 100: 104182, 2021
[4] E.J. Strait. Rev. Sci. Instrum., 77(023502), 2006
[5] G. Dong, K. G. Felker, A. Svyatkovskiy, W. Tang, J. Kates-Harbeck, Fully convolutional spatio-temporal models for representation learning in plasma science, Journal of Machine Learning for Modeling and Computing 2 (1), 2021
[6] J. L. Barr, et al., "Control Solutions Supporting Disruption Free Operation on DIII-D and EAST," invited presentation (virtual), 2020 ITER Technical Meeting (ITM) on Disruption & Mitigation, July 20-23 (2020).
[7] Schuller, F. Disruptions in tokamaks. Plasma Phys. Contr. Fusion 37, A135 (1995)
[8] General Atomics, DIII-D, accessed January 31, 2021, from http://www.ga.com/diii-d, 2021.
[9] EUROfusion ConsortiumResearch Institutions, JET: EUROfusion's Flagship Device, accessed January 31, 2021, https://www.euro-fusion.org/devices/jet/, 2014.
[10] Lehnen, M. et al. Disruptions in ITER and strategies for their control and mitigation. J. Nucl. Mater. 463, 39–48 (2015)
[11] Wroblewski, D., Jahns, G. & Leuer, J. Tokamak disruption alarm based on a neural network model of the high-beta limit. Nucl. Fusion 37, 725 (1997)
[12] Titan: advancing the era of accelerated computing. Oak Ridge National Laboratory, https://www.olcf.ornl.gov/olcf-resources/compute-systems/titan/ (accessed 24 March 2021).
[13] Traverse. Princeton Research Computing, https://researchcomputing.princeton.edu/systems/traverse accessed 24 March 2021).
[14] M.Margo et al, Current State of DIII-D Plasma Control System, Fusion Engineering and Design, Volume 150, 111368m 2929

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Fusion Energy Sciences, using the DIII-D National Fusion Facility, a DOE Office of Science user facility, under Awards DE-FC02-04ER54698; DE-AC02-09CH11466.