Received: 20 June 2022 | Revised: 19 March 2023 | Accepted: 2 May 2023 | Published on: 28 June 2023

DOI: 10.1002/ctpp.202200095

ORIGINAL ARTICLE

Contributions to Plasma Physics

Implementation of AI/DEEP learning disruption predictor into a plasma control system

William Tang[1] | **Ge Dong**[1] | **Jayson Barr**[2] | **Keith Erickson**[1] | **Rory Conlin**[1] | **Dan Boyer**[1] | **Julian Kates-Harbeck**[1] | **Kyle Felker**[1] | **Cristina Rea**[3] | **Nikolas Logan**[4] | **Alexey Svyatkovskiy**[1] | **Eliot Feibush**[1] | **Joseph Abbatte**[1] | **Mitchell Clement**[1] | **Brian Grierson**[1] | **Raffi Nazikian**[1] | **Zhihong Lin**[5] | **David Eldon**[2] | **Auna Moser**[2] | **Mikhail Maslov**[6]

[1]Princeton Plasma Physics Laboratory, Princeton, New Jersey, USA

[2]General Atomics, San Diego, California, USA

[3]Massachusetts Institute of Technology, PSFC, Cambridge, Massachusetts, USA

[4]Lawrence Livermore National Lab, Livermore, California, USA

[5]University of California Irvine, Irvine, California, USA

[6]EUROfusion Consortium, JET, Culham Science Centre, Abingdon, UK

Correspondence

William Tang, Princeton Plasma Physics Laboratory, Princeton University, PPPL, Princeton, NJ 08543, USA.
Email: wtang@princeton.edu

Funding information

Princeton Plasma Physics Laboratory

Abstract

This paper reports on advances in the state-of-the-art deep learning disruption prediction models based on the Fusion Recurrent Neural Network (FRNN) originally introduced in a 2019 NATURE publication [https://doi.org/10.1038/s41586-019-1116-4]. In particular, the predictor now features 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 adds valuable physics interpretability for the deep learning model and can provide helpful guidance for control actuators now implemented into a modern plasma control system (PCS). The advance is a significant step forward in moving from modern deep learning disruption prediction to real-time control and brings novel AI-enabled capabilities relevant for application to the future burning plasma ITER system. Our analyses use large amounts of data from JET and DIII-D vetted in the earlier NATURE publication. In addition to “when” a shot is predicted to disrupt, this paper addresses reasons “why” by carrying out sensitivity studies. FRNN is accordingly extended to use more channels of information, including measured DIII-D signals such as (i) the “n1rms” signal 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 control; and (iii) “q-min”—the minimum value of the safety factor relevant to the key physics of kink modes. The additional channels and interpretability features expand the ability of the deep learning FRNN software to provide information about disruption subcategories as well as more precise and direct guidance for the actuators in a PCS.

KEYWORDS

artificial intelligence +, Fusion Energy Science +, machine learning +, tokamak disruption prediction & control

1 | INTRODUCTION

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 timescale.[1] 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[2] and JET,[3] and future tokamak devices such as ITER.[4,5] In particular, statistical machine learning based predictors relevant to disruptions have become increasingly prevalent in recent years.[6,7] It is particularly noteworthy that in addressing this long-standing challenge, neural networks, which were considered for decades,[8] have recently taken a dramatic step forward with the advent of more powerful artificial intelligence approaches enabled by the rapid advances in high-performance computing technology at major supercomputing centers. For example, deep learning models based on the long-short term memory (LSTM) recurrent neural network (RNN)[9] and temporal convolutional neural networks (TCN)[10] have achieved breakthrough results for cross-machine predictions with the aid of leadership-class high-performance-computing (HPC) facilities.[11]

The state-of-the-art deep learning disruption prediction models based on the Fusion Recurrent Neural Network (FRNN)[9] 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” to indicate the underlying reasons/physics subcategories[12] for the imminent disruption. As an indicator of possible causes for future disruptions, the “sensitivity score” can contribute valuable physics-based interpretability for the deep learning model results, and more importantly, provide targeted guidance for the 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.[5] Key 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 Reference [9] These statistical sensitivity studies help address and answer in addition to “when” a shot is going to disrupt, some compelling reasons associated with disruption physics subcategories to help explain “why” it disrupts.

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 using modern programming languages (i.e., Python) to implement the “Keras” algorithmic scheme [Note: The book by F. Chollet provides a Python Framework using high-level Keras library with Google Tensorflow backend. Francois Chollet, “Deep Learning with Python,” 384 pages, November (2017), https://www.manning.com/books/deep-learning-with-python]. A “Keras2c” converter[13] was developed to enable integration of the AI-based predictor into the DIII-D plasma control system since the PCS for most tokamaks is written in much older C-language. Associated details are explained later in Section 2 of this paper. It is important to keep in mind that Reference [9] introduced 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 featured 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). Deploying the FRNN inference engine in a real-time PCS 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 more channels of information. For example, DIII-D signals that are known to be relevant physics-based channels include: (i) “n1rms”—a signal correlated with $n = 1$ modes with finite frequency (where n is the toroidal mode number), including the neoclassical tearing modes (NTM’s) and sawteeth dynamics; (ii) “q-min”—the minimum value of the safety factor directly relevant to important physics such as the kink modes; and also (iii) bolometer data reflecting the impurity content of the plasma. These considerations 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 (e.g., decision trees) targeting these signals for plasma condition and disruption control comes from the fact that neural nets can capture significantly greater complexity with associated predictive accuracy enabled by AI/deep learning models set up for carrying out supercomputing-enabled hyperparameter tuning enhancements of statistical accuracy for complex physical systems with large 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. In general, valuable physics-based interpretability for the deep learning model results can contribute targeted guidance for the actuators when implemented into a 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 familiar MHD safety factor at the radial location near the plasma periphery (q-95) can represent sensitive channels contributing to physics reasons associated with disruption prediction for specific cases of interest.

While significant variability can of course arise when considering general trends in large overall databases, it is important to keep in mind the potential noted earlier for significant improvement of neural nets over existing traditional algorithms. This is associated with their higher complexity whose statistical predictive accuracy can be enhanced by the use of modern HPC-assisted hyperparameter tuning with associated training carried out on path-to-exascale supercomputers at leading facilities such as ANL, LBNL, and ORNL in the United States.[11]

Overall, this paper illustrates that when more physics-related channels are statistically 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 Figure 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 actuators. Moreover, another important advantage of DL-enabled predictive capabilities is the ability to carry out forecasts significantly earlier in the evolution of the plasma state under consideration. For example, with the information included and the minimum prediction-time threshold set to 30 ms, an improved average alarm time of 100 ms was achieved for FRNN disruption prediction. This positive trend in lead time could be further tested in future real-time PCS studies in exploring strategies for disruption avoidance.

This paper also highlights the first results of the implementation of an 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 shown to be compatible with the PCS requirements. This demonstrates FRNN deep learning models are capable of disruption prediction tasks potentially useful for control efforts. We also highlight here some recent offline 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." This is useful for the interpretation of FRNN deep learning model disruption prediction results, as shown in Figure 3.

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, for example, the physics-related "n1rms" signal is included as input; 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 | IMPLEMENTATION OF FRNN DEEP LEARNING BASED MODEL INTO DIII-D PCS

The DIII-D PCS[14] is a comprehensive software/hardware system used for real-time data acquisition and feedback control of numerous actuators on the DIII-D tokamak. It regulates many plasma characteristics including shape, position, specific diverter operations, and core performance with a platform for incorporating new control algorithms that are verified and validated. For example, a disruption predictor using the shallow machine learning "random forest" method has recently been implemented and tested.[6,7] As a new category of such algorithms, the AI/deep learning FRNN software has now been integrated into this PCS and has demonstrated successful operation for a significant number of DIII-D shots in the past year (as illustrated for example in Figures 2–5 in this paper). This implementation consists of four parts: (i) pre-shot configuration; (ii) real-time data collection; (iii) processing through a Keras2c interpreter (see Section 2 below); and (iv) associated collection of the Keras2c results with associated documentation. 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 comprised of a list of normalizing factors applied to each input (as shown specifically in Reference [9]). These factors are set before the shot and applied during the real-time data collection phase. This process combines heterogenous data from multiple sources during each real-time cycle that includes diagnostics, sensor measurements, and internal calculations from other algorithms.

The data flow of the FRNN software utilize the normalized measured temporal 0D (magnitude-only) and 1D (spatial) signals as inputs. In Reference [9] the normalization scheme is discussed in detail, and this information is important to keep in mind for the interpretation of the subsequent sensitivity analysis. As explained there, all inputs are normalized to the same order of magnitude, the variance of the input quantities is appropriately scaled, and the values submitted during real-time are correlated with those trained offline. In particular, the 1D inputs are processed by a set of convolutional neural nets and then concatenated with the 0D inputs to form the input features for the long short-term memory (LSTM) network as well as for the temporal convolutional neural (TCN) network discussed in Reference [10] The outputs of the LSTM and the TCN in these studies provide complementary benchmarks for the disruption score indicating proximity of the coming disruption event.

In general, to account for significant control room adjustments such as possible recalibration of a particular diagnostic, associated modifications can include the application of pre-shot normalizing factors to match the offline functions used when training the model. It is important to note here that the values submitted during real-time correlate to values trained offline. Finally, the collection of data inputs from the experiment is inserted into a pre-defined Keras2c input data format for use in the Keras processor that is shared by multiple algorithms. The Keras processor produces a predictive result, which is then stored post-shot and provided for real-time control. Prior to the experiment, 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 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 enables avoiding the need to either use large nondeterministic software libraries or to code the entire network by hand, and thereby avoiding generating code that proves to be difficult to verify and maintain.[13]

At this point, it is useful to highlight the current progress to 2022 of the FRNN software that has successfully been run on the DIII-D PCS, including establishing a specific category to avoid potential conflicts with other algorithms. It has now demonstrated reproducible timing for representative real-time shots in the 1 ms time range. 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 in order to keep pace with attractive emerging technological advances, this AI/DL project has recently leveraged engagement with industry (NVIDIA and Concurrent) to build a unique system that enhances conventional CPU’s (central processing units) operations with modern GPU’s (graphics processing units). For example, integration of the most advanced NVIDIA A100 GPU is now completing the hardware testing phase. In future application studies, this can be expected to enable quicker and more efficient examination of the benefits of deploying our current as well as possible new AI/DL algorithms.

3 | PHYSICS-BASED SIGNALS FOR IMPROVING DISRUPTION PREDICTION

Including physics-based signals as inputs to our deep learning based models are found to improve predictive capabilities. For example, the finite frequency $n = 1$ mode amplitude is an important physical quantity, where n is the toroidal mode number. As a post-processed signal, it cannot be used for online predictions. We note though that while the “n1rms”—as well as “n2rms” and “n3rms”—signals are post-processed noncausal data, “n1rms” can help demonstrate the usefulness of including such instability-related signals in deep learning based models by shifting the “n1rms” signal input in time by 20 ms and using this value to prevent the model from seeing any future information about the plasma. This can be viewed as a useful example to explore how properties of important plasma instabilities, including the kink-like modes and neoclassical tearing modes, can eventually “lock” to the inner wall of the device and lead to disruptions. In general, the n1rms is the well-defined DIII-D signal point name that represents the $n = 1$ finite frequency magnetic perturbations.

Table 1 provides an explanation for which hyperparameters have been analyzed as we tune them for the FRNN model with the representative hyperparameter symbol and associated “well-performing” values. As shown in Figure 1, the ROC curve is generated by varying the alarm threshold.

When the $n = 1$ finite frequency mode amplitude is included as an input channel in FRNN, the disruption prediction results can be improved at both the low false-positive rate regime and the high false-positive regime, as shown in Figure 1. Quantitatively, the alarm thresholds indicated by the solid dots in Figure 1 show that the model trained with $n = 1$ finite frequency mode signal raises disruption alarms earlier than the model trained without the $n = 1$ finite frequency signal. We performed hyperparameter tuning, as introduced in,[9] for FRNN models that are trained with and without the “n1rms”

TABLE 1 Hyperparameters tuned for the FRNN model, explanations of the hyperparameter symbol, and representative well-performing values.

Hyperparameter	Explanation	Representative value
η	Learning rate	9.08 e−5
γ	Learning rate decay per epoch	0.99
N_{batch}	Training batch size	
$T_{warning}$	Warning time for target function, which becomes positive at $T_{warning}$	20
Target	Type of target function	Ttd
Model Type	Type of model (TCN/LSTM)	TCN
Nlstm	Number of LSTM layers	2
Mlstm	Number of LSTM cell per layer	128
Nt	Number of causal temporal convolutional layers	8
Ns	Number of spatial convolutional layers	2
λ	Weighting factor for positive examples	16
K_t	Size of temporal convolutional filters	11
K_s	Size of spatial convolutional filters	7
N_{Tstack}	Number of stacks of temporal convolutional blocks	2
n_{tf}	Number of temporal convolutional filters	60
n_{sf}	Number of spatial convolutional filters	20
Dropout	Dropout probability	0.05

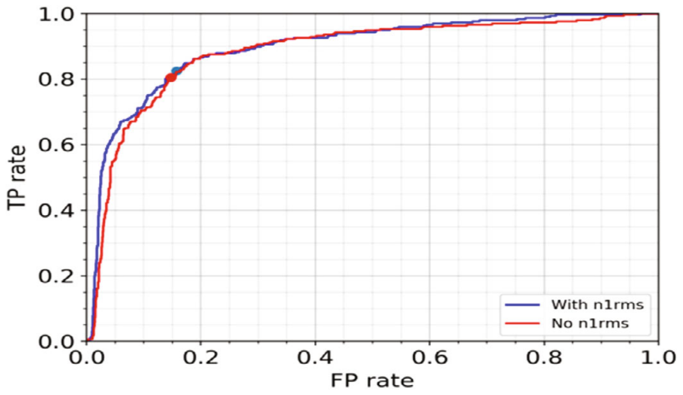

FIGURE 1 Comparison of the ROC curves with and without the $n = 1$ finite frequency mode amplitude ("n1rms"). The vertical (TP/"true-positive" rate) or "sensitivity" is plotted versus the horizontal (FP/"false-positive rate) or "probability of false alarm" at various threshold settings.

signal and reported the performance of the models that achieved the highest area under the so-called ROC curve on the validation set respectively. The list of hyperparameters tuned is shown in Table 1. A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true-positive rate (TPR) against the false-positive rate (FPR) at various threshold settings. The TPR is also known as sensitivity or probability of detection in machine learning. The false-positive rate is also known as probability of false alarm.

More importantly, as noted earlier in this discussion (and highlighted by the solid dots in Figure 1), 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 lead time are increased by more than 100 ms. To demonstrate that neural networks can effectively support the transfer of information from the $n = 1$ finite frequency mode signal and thus

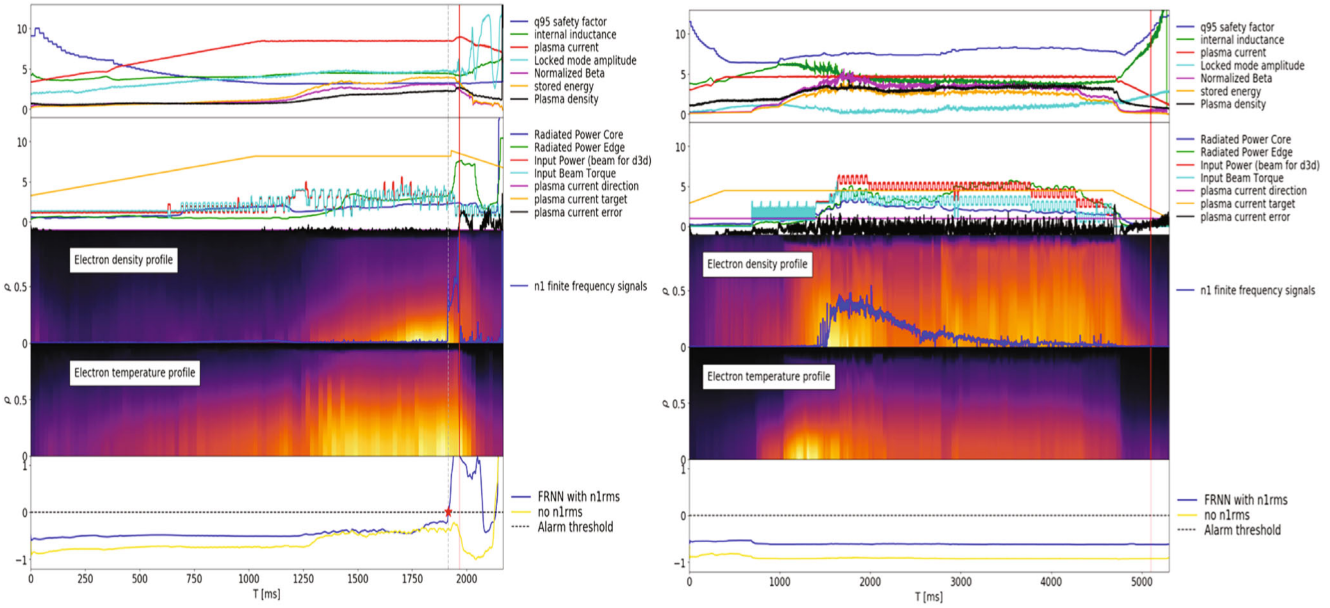

FIGURE 2 DIII-D shot number 161362 in the left panel (example of a disruptive case) and DIII-D shot number 170239 (example of a non-disruptive case) in the right panel. In each panel, the upper four sub-panels show measured signals as FRNN input, and the bottom sub-panel shows FRNN model outputs.

help provide earlier disruption alarm, an example shot from DIII-D (shot #161362) is shown in the left panel in Figure 2. At around 1.9 s, the FRNN model with "n1rms" as an input channel raised the disruption alarm following the onset of the $n = 1$ mode. Before 2 s, the "n1rms" signal diminishes while the locked mode amplitude rises. 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.

Associated studies of ROC curves using "q-min" and bolometer signals have not been adequately completed at this time due to limited data availability issues. For example, while a large portion of the shots studied did have strong q95 signals (with q95 defined as the safety factor near the plasma edge), they had very little q-min (minimum q) and q-profile information, thereby leading to a much smaller statistical database to analyze. This was also the case for the limited bolometer data, which also had associated complexity complications with proper normalizations. We comment at this point that the real-time bolometer/q data investigated did not show improvement in our disruption forecasting, and that possible future studies with larger available databases would require retraining associated models to assess whether improved performance might be achieved.

In any case, we note that 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 number 170239, as shown in the right panel of Figure 2, although a tearing mode appears at 2–3 s, the FRNN output remains at a constant low level. These results demonstrate the application of neural networks to enable discriminating between those tearing modes that can be linked to the onset of disruptions (e.g., DIII-D shot number 161362) and those that are not observed to do so (e.g., DIII-D shot number 170239).

Overall, it has been well established from many experimental observations in tokamaks (e.g., Reference [12]) that NTM activity is an important disruption, and this is in fact a key motivation for the studies in this section of our paper. It is important to keep in mind though that there have been many DIII-D shots observed where the NTM islands become large and even "locked"—but where disruption is actually avoided. Tearing modes and locked modes are in fact known to have very complicated dynamical behavior that closely relates to the disruptivity of the plasma. So, an actual specific "threshold" for NTM is quite difficult to define. We have accordingly introduced in these studies systematic AI/deep learning statistical methodology that incorporates $n = 1$ mode activity—instead of a simplistic ill-defined "locked mode threshold" to help forecast disruptions. More specifically, compared with first principles and experimental methods, it should be kept in mind that machine learning/deep learning approaches represent statistical methodology for modeling and analyzing systems that can be properly observed/measured.

4 | REAL-TIME SENSITIVITY STUDY CAPABILITY

To interpret the disruption predictive capabilities of the DL-based model, we have developed sensitivity study schemes for individual test shots—schema that can be implemented in real time along with the regular FRNN model inference engine as introduced in Section 2. As noted, the normalization of input quantities is quite important in that for each shot, the sensitivity study helps address how the neural network outputs a high disruption score and raises an associated disruption alarm at a given time. For the representative case studies in this paper (e.g., as shown in Figure 2), the input quantities considered have similar orders of magnitude and the associated variances are in line with the hyperparameters shown in Table 1 that we tuned for the FRNN model. In general, since correlations among input quantities are not taken into account, "steering" the quantity yielding the largest sensitivity score would not necessarily improve disruption avoidance. However, the results from the sensitivity study scheme can provide detailed indications of which physical quantities provide relevant "proximity guidance" for disruptive scenarios, and this information may directly aid control efforts by identifying the appropriate actuator to possibly optimize "proximity to a disruption." This aspect is further noted in comments in Section 5.0 and is currently being pursued in collaboration with Barr et al.[15] in future AI/DL FRNN control investigations that include the deployment of actuators. Finally, it should be kept in mind that while sensitivity studies specific to individual shots can be carried out in real time, the original Nature paper[9] focused only on the general training and test set without any real-time application. In addition, the original RNN workflow renders a Python model that cannot be directly used in a PCS.

4.1 | Calculation of the sensitivity score

In Reference [9] 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 nondisruptive shots. For noise-aware models that are trained with dropped-out signals,[8] and are 'familiar' with all-zeroes-signals, we suppress a signal by directly replacing it with zeros. While the outputs of these two schemes do not of course encompass all possibilities, they are generally qualitatively consistent and illustrate the robustness and promise of these sensitivity study results with respect to the replacement values.

An example of the sensitivity study result of DIII-D shot number164582, Figure 3 illustrates where the evolutions of the sensitivity scores of each signal are plotted as a function of time. At around 2–2.7 s, the disruption score slowly rises up, as shown in the last panel in Figure 4, 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 Figure 4. The influx of impurities is followed by the rise of the core radiation, which leads to a disruption alarm at around 3 s. At the disruption alarm time, the channels with the highest sensitivity scores are core radiation and q-95. Such sensitivity studies could accordingly be used in future studies to suggest PCS operations, for example, that might engage ECH in the core region to flush impurities or to simply raise q-95. We also note here that at around 3.5 s, 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 Figure 3. Around this time, every channel becomes sensitive, and the sensitivity scores begin to change rapidly as the plasma current error rises.

4.2 | Sensitivity score from "Zero-value replacement procedure" for "noise-aware" models

We show the sensitivity study result for DIII-D shot number162975 in Figure 5, using the "zero-value replacement procedure" for individual channels during inference studies using noise-aware models. It is illustrated that after 1.5 seconds,

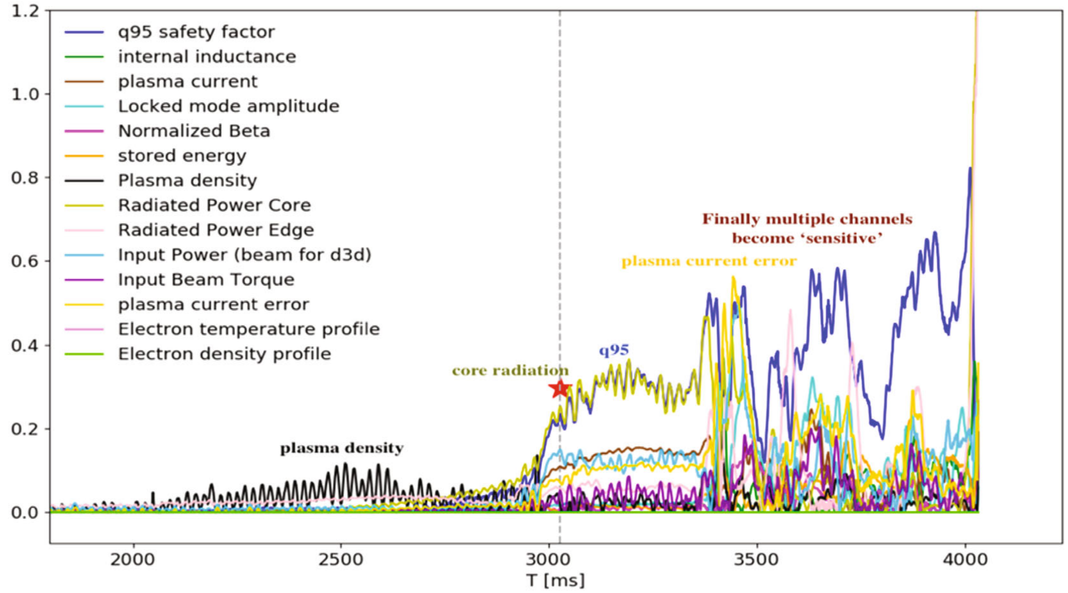

FIGURE 3 Illustration of the contributions of physics-based signals to the temporal evolution of the sensitivity score (dimensionless) for disruptions that are plotted on the *y*-axis for the representative DIII-D shot number 164582.

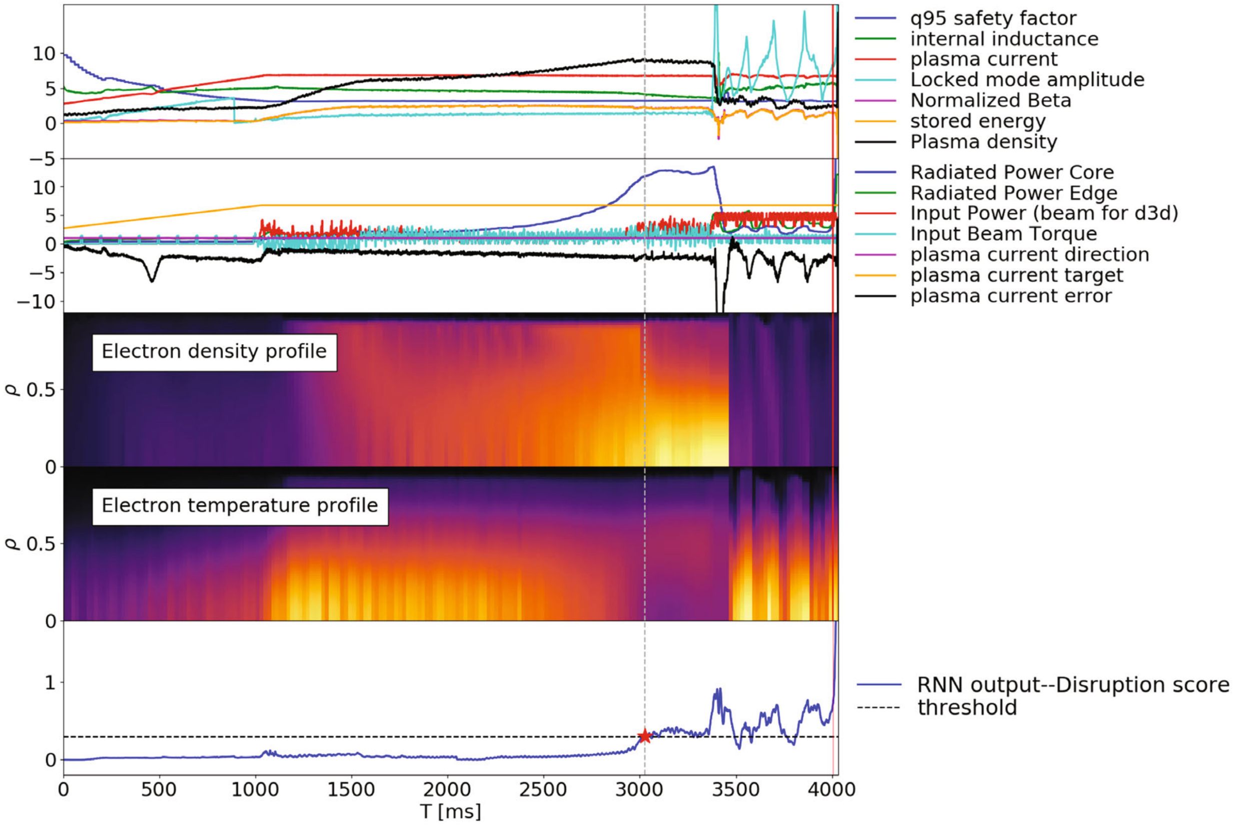

FIGURE 4 Evolution of each normalized physical signal for DIII-D shot number 164582 in the upper four panels. The bottom panel shows the time history of the FRNN output.

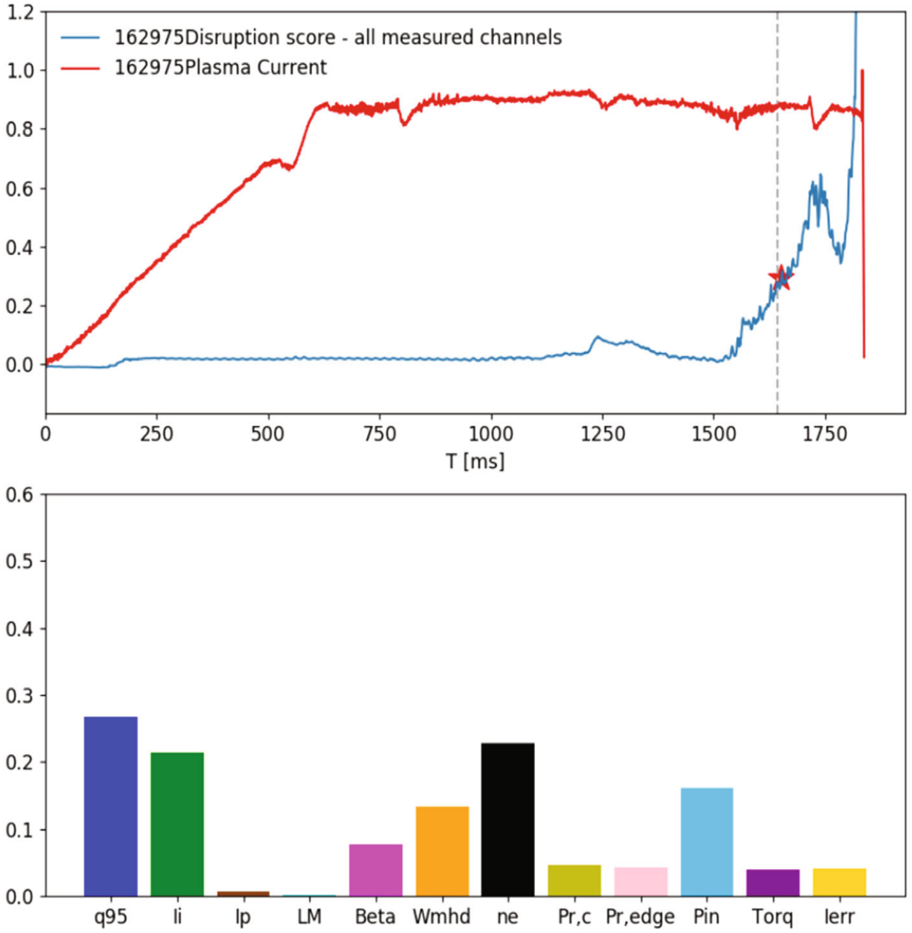

FIGURE 5 DIII-D shot number 162975. The upper panel shows the evolution of the plasma current (red line) and the FRNN output of the disruption score (blue line—with the lower panel showing the sensitivity scores (for associated signals such as q95, etc.) at the time of the disruption alarm (red star in the upper panel).

the disruption score begins to increase and raises 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. Sensitive channels including internal conductivity, the q95 safety factor, and plasma density all can contribute to a general deterioration of plasma shape and plasma control. This interpretation is qualitatively consistent with the observation of experimental characteristics. For example, at around 1.3 s, the plasma shape begins to alter as the X point moves off-target in this shot. Future collaborative studies with EFIT[13] can help to ascertain the magnitude of such changes. In Figure 5, a tearing mode begins to develop at 1.45 s—which then plausibly evolves into the locked mode at round 1.6–1.7 s when the disruption alarm rises. 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 that lock are important causes for the disruptions, the actual locked mode amplitude is not a sensitive signal for both of these shots. 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. This is supported by the fact that we did not use the "n1rms" signals as input for these studies. Accordingly, such sensitivity studies are indeed potentially capable of contributing significantly to real-time disruption mitigation and avoidance investigations.

5 | SUMMARY AND ASSOCIATED FUTURE INVESTIGATIONS

In the present paper, we have described the implementation of our AI/deep learning disruption prediction software into the DIII-D plasma control system (PCS). This includes the introduction of a new method for interpreting results from deep neural networks using a sensitivity study methodology with significant implications for real-time actionable integration into future plasma control systems. Key conclusions and consequences from demonstrating the sensitivity of the alarm from the AI/deep learning FRNN disruption predictor include illustrating the interpretability of deep learning based models and capability to provide physics-based information for the PCS when a disruption alarm is raised. The present studies are indicative that if more signals can be included in the training database, it can be expected that better predictive capability in associated future investigations will likely emerge. Here we are highlighting the initial exciting ability of the neural network to discriminate between disruptive and nondisruptive tearing modes. In the future, additional useful guidance for the PCS could be generated, for example, by carrying out parallel inference studies with small variances for key quantities displayed in Figure 5 (density, li, …) with results leading to a possible steepest-descent trend in the disruption score that might provide an alternative approach to the current sensitivity measure. Overall, in planned ongoing and future investigations, we will extend and interconnect stimulating features of this work by providing the deep learning sensitivity output in real time into the proximity control architecture designed for handling major disruption causes in the DIII-D PCS.[15]

ACKNOWLEDGMENTS

The authors are grateful to Dr. David Humphreys of General Atomics, DIII-D for his careful review with helpful suggestions for improvement and clarification that have been integrated into this manuscript. Material for these studies 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-SC003444, DE-FC02-04ER54698; DE-AC02-09CH11466, DE-AC52-07NA27344, DE-SC0020337, DE-SC0014264. This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training program 2014–2018 and 2019–2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission. Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the US Government or any agency thereof.

DATA AVAILABILITY STATEMENT

Data available on request from the authors.

ORCID

Jayson Barr https://orcid.org/0000-0001-7768-5931
Cristina Rea https://orcid.org/0000-0002-9948-2649

REFERENCES

[1] F. Schuller, *Plasma Phys. Contr. Fusion* **1995**, *37*, A135.

[2] W.M. Solomon for the DIII-D Team, *Nucl. Fusion* **2017**, *57*, 102018.

[3] EUROfusion Consortium Research Institutions. JET: EUROfusion's Flagship Device. **2014**. https://www.euro-fusion.org/devices/jet/

[4] M. Lehnen, K. Aleynikova, P. B. Aleynikov, D. J. Campbell, P. Drewelow, N. W. Eidietis, Y. Gasparyan, R. S. Granetz, Y. Gribov, N. Hartmann, E. M. Hollmann, V. A. Izzo, S. Jachmich, S. H. Kim, M. Kočan, H. R. Koslowski, D. Kovalenko, U. Kruezi, A. Loarte, S. Maruyama, G. F. Matthews, P. B. Parks, G. Pautasso, R. A. Pitts, C. Reux, V. Riccardo, R. Roccella, J. A. Snipes, A. J. Thornton, P. C. de Vries, *J. Nucl. Mater.* **2015**, *463*, 39.

[5] E. J. Strait, J. L. Barr, M. Baruzzo, J. W. Berkery, R. J. Buttery, P. C. de Vries, N. W. Eidietis, R. S. Granetz, J. M. Hanson, C. T. Holcomb, D. A. Humphreys, J. H. Kim, E. Kolemen, M. Kong, M. J. Lanctot, M. Lehnen, E. Lerche, N. C. Logan, M. Maraschek, M. Okabayashi, J. K. Park, A. Pau, G. Pautasso, F. M. Poli, C. Rea, S. A. Sabbagh, O. Sauter, E. Schuster, U. A. Sheikh, C. Sozzi, F. Turco, A. D. Turnbull, Z. R. Wang, W. P. Wehner, L. Zeng, *Nucl. Fusion* **2019**, *59*, 112012.

[6] C. Rea, R. S. Granetz, K. Montes, R. A. Tinguely, N. Eidietis, J. M. Hanson, B. Sammuli, *Plasma Phys. Control. Fusion* **2018**, *60*, 084004.

[7] K. J. Montes, C. Rea, R. A. Tinguely, R. Sweeney, J. Zhu, R. S. Granetz, *Nucl. Fusion* **2019**, *59*, 096015; K. J. Montes, et al., *Nucl. Fusion* **2021**, *61*; C. Rea, K. J. Montes, A. Pau, R. S. Granetz, O. Sauter, *Fusion Sci. Technol.* **2020**, *76*, 912.

[8] D. Wroblewski, G. Jahns, J. Leuer, *Nucl. Fusion* **1997**, *37*, 725.

[9] J. Kates-Harbeck, A. Svyatkovskiy, W. Tang, *Nature* **2019**, *568*, 526.

[10] G. Dong, K. G. Felker, A. Svyatkovskiy, W. Tang, J. Kates-Harbeck, *J. Mach. Learn. Modeling Comput.* **2021**, *2*, 49.

[11] Summit. www.olcf.ornl.gov **2021**.

[12] P. C. de Vries, M. F. Johnson, B. Alper, P. Buratti, T. C. Hender, H. R. Koslowski, V. Riccardo, *Nucl. Fusion* **2011**, *51*, 053018.

[13] R. Conlin, K. Erickson, J. Abbate, E. Kolemen, *Eng. Appl. Artif. Intell.* **2021**, *100*, 104182.

[14] M. Margo, B. Penaflor, H. Shen, J. Ferron, D. Piglowski, P. Nguyen, J. Rauch, M. Clement, A. Battey, C. Rea, *Fusion Eng. Des.* **2020**, *150*, 111368.

[15] 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**.

How to cite this article: W. Tang, G. Dong, J. Barr, K. Erickson, R. Conlin, D. Boyer, J. Kates-Harbeck, K. Felker, C. Rea, N. Logan, A. Svyatkovskiy, E. Feibush, J. Abbatte, M. Clement, B. Grierson, R. Nazikian, Z. Lin, D. Eldon, A. Moser, M. Maslov, *Contrib. Plasma Phys.* **2023**, *63*(5-6), e202200095. https://doi.org/10.1002/ctpp.202200095