

1 Fully Convolutional Spatio-Temporal Models for Representation Learning in
2 Plasma Science

3
4
5 Ge Dong^{1(a)}, Kyle Gerard Felker², Alexey Svyatkovskiy³, William Tang^{1*}, Julian Kates-
6 Harbeck^{1*}

7
8 ¹ Princeton Plasma Physics Laboratory, Princeton, NJ

9 ² Argonne National Laboratory, Lemont, IL

10 ³ Microsoft Corporation, Redmond, WA

11 *joint supervision

12 ^(a)gdong@princeton.edu

13
14 Abstract:

15
16
17 We have trained a fully convolutional spatio-temporal model for fast and accurate representation
18 learning in the challenging exemplar application area of fusion energy plasma science. The onset
19 of major disruptions is a critically important fusion energy science issue that must be resolved
20 for advanced tokamak plasmas such as the \$25B burning plasma international ITER experiment.
21 While a variety of statistical methods have been used to address the problem of tokamak
22 disruption prediction and control, recent approaches based on deep learning have proven
23 particularly compelling. In the present paper, we introduce further improvements to the fusion
24 recurrent neural network (FRNN) software suite, which delivered cross-machine disruption
25 predictions with unprecedented accuracy using a large database of experimental signals from two
26 major tokamaks. Up to now, FRNN was based on the long short-term memory (LSTM) variant
27 of recurrent neural networks to leverage the temporal information in the data. Here, we
28 implement and apply the “temporal convolutional neural network (TCN)” architecture to the
29 time-dependent input signals. This allows highly optimized convolution operations to carry the
30 majority of the computational load of training, thus enabling a reduction in training time, and the
31 effective use of high performance computing (HPC) resources for hyperparameter tuning. At the
32 same time, the TCN based architecture achieves better predictive performance when compared
33 with the LSTM architecture for various tasks for a representative fusion database.
34
35
36
37
38
39
40
41
42
43
44
45

46 I. Introduction

47

48 Deep learning has become an increasingly important methodology for the effective analysis and
49 interpretation of big data in modern social and scientific areas (Webb, 2018; Schmidt et al.,
50 2019; Reichstein et al., 2019). In this study, we discuss the application of deep learning models
51 in the prominent exemplar problem of disruption predictions in tokamaks (Kates-Harbeck et al.,
52 2019), which are magnetic fusion experimental devices with large numbers of advanced
53 diagnostics to monitor spatio-temporal plasma performance.

54

55 In many toroidal plasma devices such as tokamaks and spherical toruses (ST's) (Gerhardt et al.,
56 2013), disruptions are observed as sudden and dangerous events that induce rapid release of
57 particles and energy to the device wall (Schuller, 1995). A typical disruption brings the plasma
58 experiment (the “shot”) to an abrupt end and, because of the associated large rapid energy
59 release, it can also seriously damage the device – especially in larger systems such as ITER
60 (Strait et al., 2019). Accordingly, the development of a plasma control system (PCS) with the
61 ability to reliably detect and subsequently mitigate or avoid the majority of the disruption events
62 (Hollmann et al., 2015) is broadly regarded as the most important milestone for establishing the
63 viability of future larger tokamak devices to deliver a fusion energy reactor.

64

65 To robustly mitigate or prevent disruptions, the first step for the PCS is to accurately predict
66 disruptions as early as possible. Traditional methods for disruption studies and predictions range
67 from using simple empirical formulae and analytic expressions to more sophisticated first-
68 principles-based simulations, such as the magnetohydrodynamic (MHD) models (Glasser and
69 Kolemen, 2018) and the gyrokinetic models (Liu et al., 2014). These simulations are used to
70 study the dynamics and mechanisms of disruptions, and accordingly to advance their prediction
71 and also their possible avoidance through active plasma control. Empirical expressions and even
72 MHD models are generally simple and fast enough to be implemented in the PCS and can
73 thereby aid real-time disruption predictions and control. However, these models often contain
74 insufficient physics information to correctly predict complex or novel scenarios, including device
75 operations for different parametric and/or hardware regimes, such as those involving, for
76 example, reactor-relevant wall materials (Matthews et al., 2011). In order to better capture
77 nonlinear dynamics and physics associated with realistic magnetic geometry, large-scale first-
78 principles-based simulations can be engaged. However, the vast computational resources
79 required to allow real-time predictive capabilities for such simulations make an implementation
80 in the PCS infeasible. Consequently, complementary to the two aforementioned categories of
81 models, emerging big-data-driven methodologies have become an increasingly powerful modern
82 approach addressing the grand challenge of prediction and control of tokamak disruptions
83 (Pautasso et al., 2002). We have accordingly developed a deep learning capability of general
84 computational science interest that is capable of enabling significant progress towards resolving
85 this major application science exemplar problem via effective utilization of the hardware
86 capabilities of modern leadership class supercomputing systems. This involves the training of
87 multiple models with distinct architectures within a single software suite adaptable to different
88 temporal and spatial learning tasks – with natural connections to enabling ensemble schemes for
89 highly accurate prediction.

90

91 We focus on and refer to models that can be easily built and extended in a layered fashion – as
92 well as optimized (trained) using automatic differentiation and back-propagation (such as the
93 neural networks) – as “deep learning” models. Machine learning models that do not have this
94 property are referred to as “classical” models. Examples of classical models include support
95 vector machines (López et al., 2012) and random forest algorithms (Rea et al., 2019), both of
96 which have been applied successfully to advance disruption prediction capability. So far, a key
97 advantage of deep learning models has been their ability to perform cross-machine predictions –
98 forecasting the plasma behavior in an experiment never seen during training and or validation.
99 This is especially key for establishing the relevance of such studies for ITER, as ITER will not
100 be able to withstand enough disruptions for a large training data set (de Vries et al., 2016).

101
102 In this paper, we introduce the implementation of a temporal convolutional neural network
103 (TCN) (Oord et al., 2016) for the temporal representation of the extensive database of fusion
104 energy science (FES) input signals of interest (Kates-Harbeck et al., 2019). The associated TCN
105 architecture based on dilated causal convolutions (Bai et al., 2018) has some advantages
106 compared with the long short-term memory (LSTM) architecture in FRNN. The TCN
107 architecture will be introduced in detail in the next section.
108

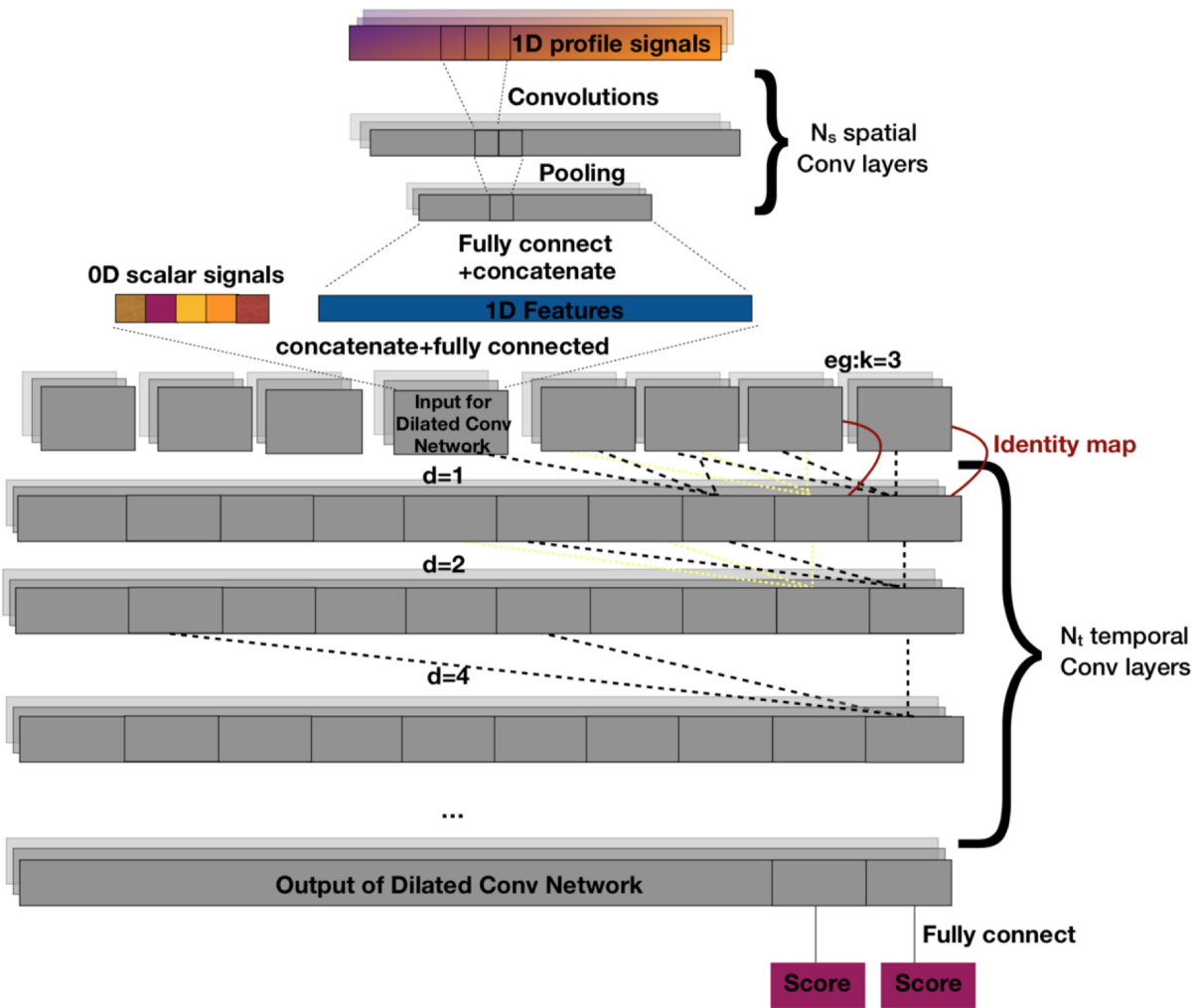
109 Comparisons between TCN and LSTM based architectures for various fusion databases are
110 described in detail in Section III of this paper. While TCN based architectures generally achieve
111 superior predictive power and improved computational performance, we note that for certain
112 tasks and databases, the original LSTM-based model can still outperform all alternatives. From a
113 general computational science perspective, this finding highlights the importance of maintaining
114 multiple architectures in a modern software suite such as FRNN. Moreover, the choice of
115 temporal processing layer (such as LSTM cell versus TCN) can be viewed as a high level
116 architectural hyperparameter for FRNN. This framing aligns with our vision for an AI/DL based
117 platform with flexible and adaptive model architectures that can be automatically
118 hyperparameter tuned for various tasks and databases associated, for example, with future fusion
119 plasma predictions and analysis tasks. This in turn has implications for building capabilities to
120 face future real-time plasma control challenges. **To increase the robustness of our models, we
121 developed noise-aware training scheme within the FRNN framework. The noise-aware training
122 can help achieve predictive capabilities of the deep learning based models during real-time
123 experiment in scenarios where certain diagnostics are unavailable or random data contamination
124 happens.**

125
126

127 II. Model Architecture

128
129 Figure 1 shows the schematic of the new deep learning architecture based on TCN’s introduced
130 here into FRNN. As in the previous LSTM based models, the TCN-based FRNN models process
131 inputs composed of two main types of signals: 0D scalar signals (such as the plasma current),
132 and 1D profile signals (such as the electron temperature profile). The descriptions and numerical
133 properties of the fourteen 0D and two 1D signals can be found in Table 1. Example time series of
134 signals available on DIII-D are shown in the top four panels of Figure 3. At each time step, the
135 1D profiles are first “spatially” processed by a sub-network consisting of N_s convolutional

136 layers. The output of this network contains a representation of the 1D features, as shown by the
 137 blue bar in Figure 1. The 1D features are then concatenated with the 0D scalar signals, and
 138 together form the complete input channels for the N_t dilated convolutional layer blocks. Each of
 139 these convolutional layer blocks consists of dilated causal convolutional layers, activation layers,
 140 optional dropout layers, normalization layers, as well as an additive identity map (which
 141 maintains the stability of the neural network when it becomes “deep”). Details of the
 142 convolutional layer blocks were introduced in (Bai et al., 2018).
 143



144
 145
 146 Fig. 1. The detailed schematic of our deep learning model based on temporal convolutional
 147 architecture. N_s is the number of convolutional layers for spatial information processing at each
 148 time step, N_t is the number of dilated temporal convolutional layers, d is the dilation factor, and k
 149 is the filter size for the dilated convolutional layers.
 150
 151
 152
 153
 154

Signal description	Numerical scale DIII-D	Numerical scale JET
Plasma current	3.8 e-1 MA	5.03 e-1 MA
Plasma current target	3.9 e-1 MA	Not available on JET
Plasma current error	3.1 e-2 MA	Not available on JET
Plasma current direction	1.0	Not available on JET
Internal Inductance	2.02 e-1	1.51 e-1
Plasma density	1.19 e19 m ⁻³	4.69 e19 m ⁻³
Input power (beam for DIII-D)	1.85 e6 W	4.47 e6 W
Radiated power core	4.58 e2 W	4.05 e4 W
Radiated power edge	4.94 e2 W	2.72 e4 W
Stored energy	2.79 e5 J	1.2 e6 J
Locked mode amplitude	1.14 e-6 T	5.72 e-5 T
Safety factor q95	1.0	1.0
Normalized plasma pressure	6.91 e-3	Not available on JET
Input beam torque	1.47 Nm	Not available on JET
Electron temperature profile	9.53 e-1 keV	1.53 keV
Electron density profile	1.47 e19 m-3	2.98 e19 m ⁻³

155 Table 1. Measured DIII-D and JET experimental signals used in FRNN
156

157 Compared with the LSTM based models, TCN based models have two main advantages for
158 processing the temporal information. First, instead of carrying over historical information in a
159 recurrent fashion, the TCN directly fetches it through a time series, and can accordingly
160 “remember” such information from a more distant past. How to effectively learn long-term
161 dependencies for predictions involving time series databases is an active area of modern
162 computational science research. While the LSTM is an improved architecture over the standard
163 recurrent neural network (RNN) – which has short memory due to the exploding and vanishing
164 gradient problems (Bengio et al., 1994) – it can still lose distant information through operations
165 on the cell state, which carries its long term memory. On the other hand, while the temporal
166 receptive field of regular convolutions grows linearly with network depth (necessitating
167 prohibitively deep architectures for long-term memory tasks), the temporal receptive field of
168 dilated convolutions grows exponentially with network depth. TCNs can thus capture long
169 distance dependencies with a modest number of layers.
170

171 The second advantage of the TCN architecture is that it is easily amenable to accelerated training
172 via model parallelism. The serial implementation of the model is straightforward; i.e., in addition
173 to fully connected layers and activation functions, only the convolution operation (applied to
174 both the spatial and temporal representations) is required. The resulting feedforward network
175 does not incorporate gated functions or recurrent connections. In contrast, the recurrent nature of
176 the LSTM and other models based on RNN are generally hard to parallelize within training
177 examples on the model level, as commented by Bai et al., 2018. We highlight the fact that for our
178 exemplar FES application of interest, the training and inference on long experimental runs with
179 dense temporal measurements leads to a large memory footprint – thereby making the TCN’s
180 efficient model parallelism a particularly attractive feature in the context of high performance
181 computing (HPC). Specifically, the TCN architecture can effectively utilize the hardware

182 capabilities of modern leadership class supercomputing systems – a fact that will be
183 demonstrated later in this paper.

184
185 With a chosen timestep of 1ms, and a typical shot duration on DIII-D of several seconds (tens of
186 seconds on JET), the typical length of a time series in our data set is thousands to tens of
187 thousands of steps. Since the effective history length of a normal convolutional layer is $k-1$,
188 where k is the convolution filter size, regular convolutional neural networks have a receptive
189 field linear in depth. By contrast, that of dilated convolutional neural networks, where each layer
190 has effective history length of $(k-1)*d$, can be exponential if the dilation factor d is grown
191 accordingly. The need for dilations (see dashed connections in Figure 1) arises if we need to
192 allow the network output near the end of the shot to depend on early plasma behavior – and to be
193 able to do so without requiring unwieldy network depth.

194
195 The output of the dilated convolutional layer blocks feeds into a final fully-connected layer that
196 combines the information from all of the hidden units. It then outputs the disruption score which
197 measures the likelihood of an imminent disruption at each time step. The definition of this
198 disruption score, or the “target” that FRNN is trained on, is effectively a hyperparameter that can
199 be tuned. A detailed explanation of the tradeoffs involved in the selection of the target function is
200 provided in the “Methods” section of Kates-Harbeck et al., 2019.

201
202 For a disruptive shot, if the disruption score rises above the pre-set “alarm threshold” before the
203 “warning time” (the latest acceptable alarm time before the actual disruption), it would count as a
204 true positive (TP) prediction. For a non-disruptive shot, if the disruption score rises above the
205 “alarm threshold” at any time, it would count as a false positive (FP) prediction. By shifting the
206 ‘alarm threshold’ from minus infinity (model predicts disruption for every shot) to infinity
207 (model predicts no disruption for any shot), a receiver operating characteristic curve (ROC
208 curve) is produced. We use the area under the curve (AUC) of the ROC curve on the test data
209 prediction results as the metric to evaluate model performance. Although for actual tokamak
210 operations, a fixed alarm threshold is required, for offline studies such as those introduced in this
211 work, the AUC of the ROC can effectively represent the general model performance.

212 213 III. Training and Prediction Results

214
215 The data for this work comes from the DIII-D tokamak located at General Atomics, San Diego,
216 CA (<http://www.ga.com/diii-d>), and the EUROfusion JET tokamak located at the Culham
217 Science Centre for Fusion Energy in the UK (<https://www.euro-fusion.org/devices/jet/>). The
218 DIII-D data is sampled from shot numbers ranging from 125500 to 168555, while the JET data
219 are sampled from the carbon wall campaigns C23-C27b and the ITER-like wall campaigns C28-
220 30. The different wall conditions in the two types of JET campaigns result in distinct plasma
221 dynamics and disruption mechanisms. In this work, time step size for the signal data is chosen to
222 be 1ms. Using the same data selection and preprocessing procedure as described in the Methods
223 section of Kates-Harbeck et al., 2019, we consider a shot to be valid within the database if and
224 only if all of the relevant signals contain data for a period of time longer than the warning time.
225 In order to assess model performance on DIII-D with a 1 second warning time, the DIII-D shots
226 are required to have non-NAN and non-flat data in all signals for at least 1 second in order to be

227 considered “valid” and to thus be added to the FRNN training/validation/testing dataset. The
 228 numbers of shots in training, validation, and testing sets as listed in parenthesis in Table 2.
 229
 230

	Single machine			Cross Machine		
	DIII-D (1702)	JET-CW(1956)		DIII-D (2268)	JET-CW(1956)	
Training (#shots)	DIII-D (837)	JET-CW(962)		DIII-D (1117)	JET-CW(962)	
Validation (#shots)	DIII-D (846)	JET-ILW(1133)		JET-ILW(1133)	DIII-D(846)	
Testing (#shots)	30ms	0.2s	1s	30 ms	30ms	30ms
Warning time	0.93	0.90	0.72	0.95	0.85	0.76
FRNN 0D-LSTM	0.93	0.90	0.74	0.95	0.91	0.73
FRNN 0D-TCN	0.93	0.89	0.80		0.84	
FRNN 1D-LSTM	0.93	0.91	0.79		0.86	
FRNN 1D-TCN						

231
 232 Table 2. Prediction results of the tuned models on the test datasets, measured as AUCs at the
 233 warning time before a disruption. Performance of the best models are highlighted in bold. Four
 234 FRNN models – trained with (1D) or without (0D) the 1D profiles, based on the LSTM or TCN
 235 architectures – are compared when trained on DIII-D and JET shots. For single machine
 236 prediction tasks using the DIII-D database, we carried out hyperparameter tuning for three
 237 different warning times, 30ms, 0.2s and 1s.
 238

239 In Table 2 we report the prediction results from four distinct FRNN architectures across several
 240 different experimental databases. Specifically, these four schemes are either LSTM based or
 241 TCN based models, each trained with and without the 1D profile information. We tuned
 242 hyperparameters for each architecture to select the best performing model for each of the 6
 243 different tasks. These six tasks are as follows: (i-iii) prediction of DIII-D disruptions with 3
 244 different warning time cutoffs using models trained on separate DIII-D data; (iv-v) predictions
 245 for JET ITER-like wall shots disruptions with 30ms warning time using models trained on JET
 246 carbon wall data, or on DIII-D data; and (vi) prediction for DIII-D disruptions with 30ms
 247 warning time using models trained on JET carbon wall data. Since the JET carbon wall datasets
 248 did not include profile information, there are no FRNN 1D results for models trained or tested
 249 with this data.
 250

251 To tune the hyperparameters of the neural network for each of the twenty entries corresponding
 252 to different predictive tasks in Table 2, we randomly selected 40 sets of hyperparameters chosen
 253 from a reasonable range of possible values, and then trained 40 models in parallel using the

254 **training and validation data.** Main hyperparameters with their representative values for the TCN
 255 based models are summarized in Table 3. In total, 800 (20 x 40) models were trained. After
 256 training all 40 models, we selected the best model based on its performance on the validation
 257 dataset. Finally, we examined the accuracy of each optimal model using the appropriate test
 258 dataset and summarized the test performance in Table 2. **The predictive performance is sensitive**
 259 **to the number of temporal and spatial convolutional filters, and the number of temporal**
 260 **convolutional blocks. When the size of the network is too small, it is not sufficient to represent**
 261 **the physical processes leading to disruptions. The predictions are only sensitive to the number of**
 262 **causal temporal convolutional layers when it is below 8, indicating that the effective memory of**
 263 **the plasma contributing to disruption is on the order of hundreds of milliseconds, consistent with**
 264 **the general plasma equilibrium evolution time scale.**

265

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 (function linear in time to disruption)
N_t	Number of causal temporal convolutional layers	8
N_s	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
$N_{T\text{stack}}$	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

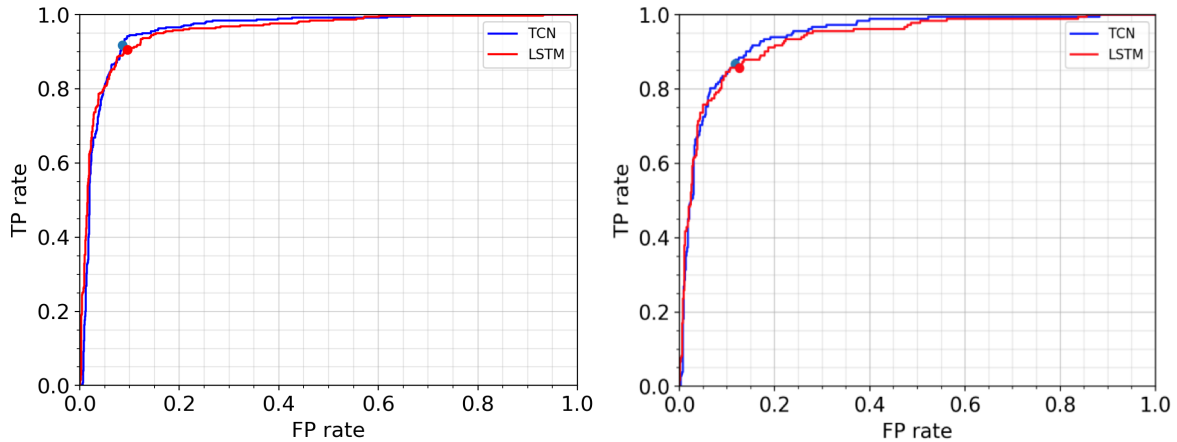
266 Table 3. Hyperparameters to be tuned for the TCN based FRNN model, explanations of the
 267 hyperparameter symbol, and representative well performing values.

268

269

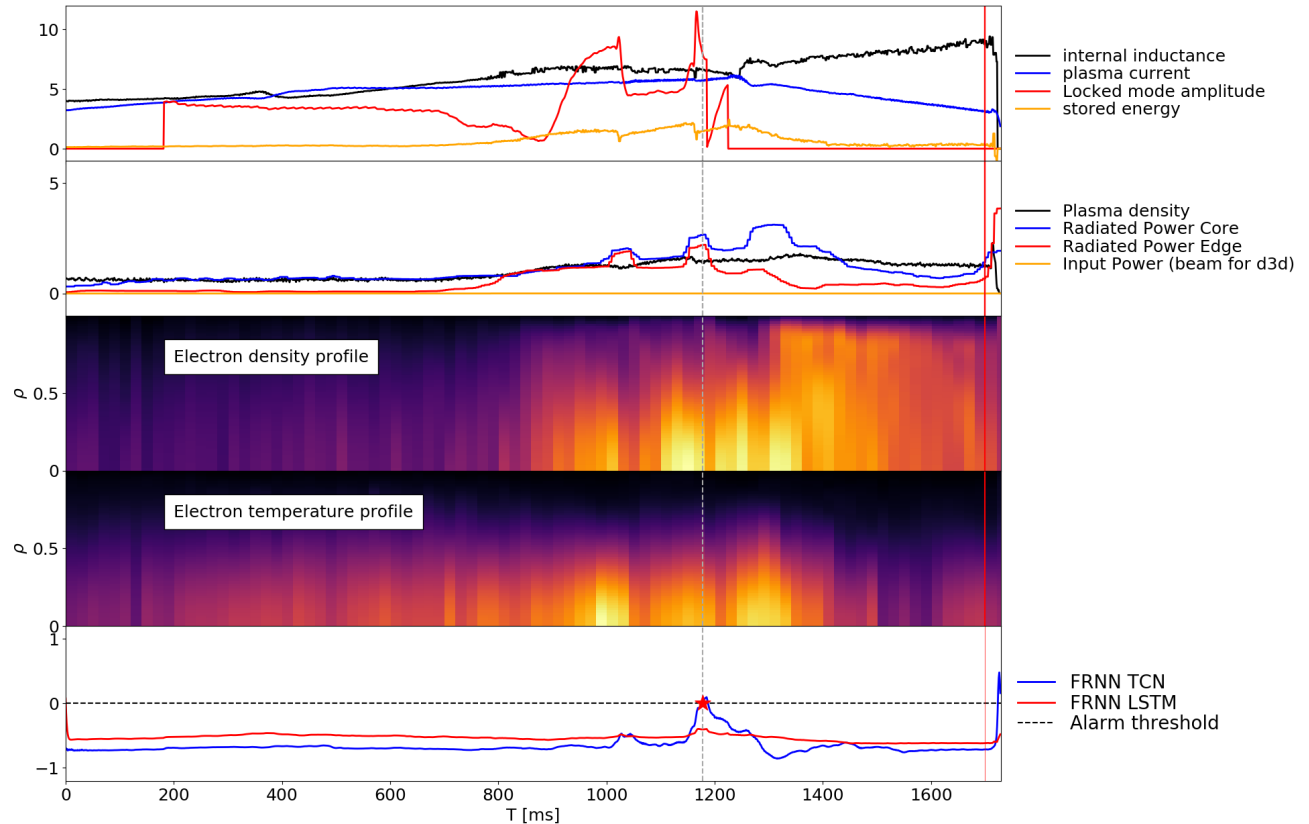
270 Comparing results from the four schemes studied, the best performing model for each task is
 271 highlighted in bold in Table 2. From this round of hyperparameter tuning run, for most of the
 272 tasks, the TCN-based model in FRNN performs equally well or better than the original LSTM-
 273 based model. As an example, the ROC curves for the best performing models (based on TCN or
 274 LSTM) on the DIII-D test dataset with 0.2 second warning time and including 1D profiles are
 275 shown in Figure 2. The two models perform equally well for the low false positive rate regime,
 276 and the TCN-based model performs slightly better in the high true positive rate regime. As the
 277 high true positive rate regime is of greater importance to disruption prediction models for future
 278 machines that could not afford false negative results, this result demonstrates an example where
 279 considering multiple deep learning models can contribute to stronger predictive power for

280 disruptions. Figure 3 shows an example DIII-D shot in the test dataset, and the output of these
281 two models. While both models show some change in their disruption score shortly before the
282 1200ms mark, only the TCN model shows a sufficient change to trigger a correct alarm. Near the
283 end of the shot, the TCN model also outputs a higher disruption score, although this alarm is
284 within the 30ms warning time, and is considered “too late”.
285



286
287
288
289
290
291
292
293

Fig 2. Comparison of ROC curves on the DIII-D training (left panel) and test (right panel) dataset with 0.2s warning time for the optimal FRNN 1D models, based on the TCN (blue) and LSTM (red) architectures. The solid dots indicate model performance at the optimal alarm threshold determined on the validation set. Both models demonstrate good generalization of the optimal alarm threshold from DIII-D validation data to DIII-D test data.



294
295

296 Fig 3. Example prediction on DIII-D shot # 147206. The top two panels show eight
 297 representative 0D scalar signal channels (normalized by numerical scale on DIII-D as listed in
 298 Table 1), and the next two panels show two 1D profile signals channels as FRNN inputs. Vertical
 299 axis of the 1D profile signals are the normalized toroidal magnetic flux ρ . The last panel shows
 300 the disruption scores returned by FRNN using TCN (blue line) and LSTM (red line) based
 301 models. The FRNN TCN model raised positive disruption alarm before 1200 ms when reaching
 302 disruption alarm threshold indicated by the horizontal dashed line. The signals are plotted up to
 303 the time of disruption. The solid vertical red line shows the latest warning time (30ms before the
 304 disruption). It is important to highlight the finding here that while both models respond
 305 noticeably around the indicated disruption alarm time, only the TCN based model correctly
 306 triggers the disruption alarm around 0.5 second before the actual disruption.

307

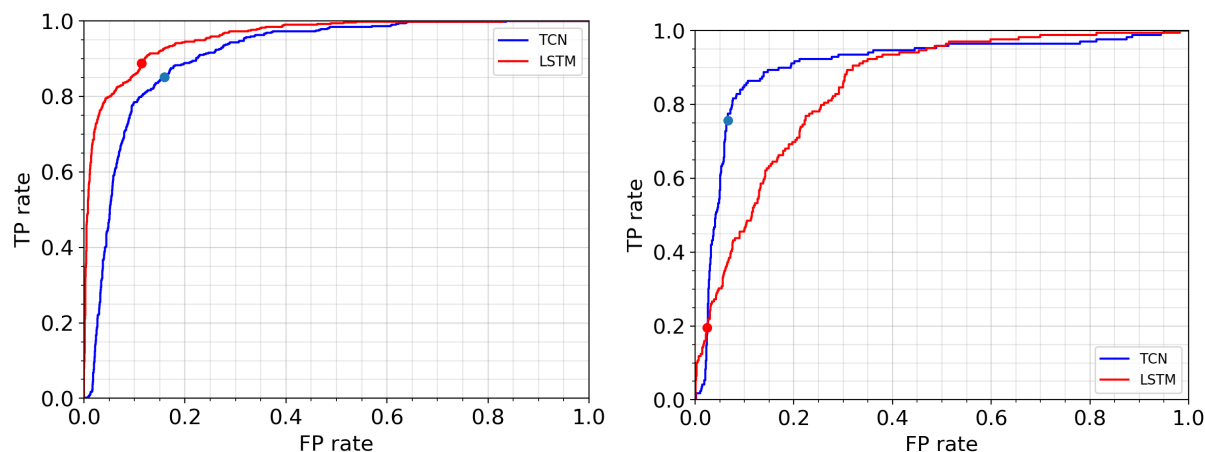
308 For the cross-machine prediction task where the model was trained only on DIII-D shots and is
 309 tasked to predict disruptions in the JET ITER-like wall shots, the TCN architecture offers
 310 significantly improved accuracy with less overfitting. The problem of overfitting is hard to
 311 overcome for this cross-machine predictive task due to drastically different physical parameters
 312 for DIII-D and JET plasmas. A comparison of the best TCN based model and the best LSTM
 313 based model performance on the cross-machine prediction task is shown in Figure 4. Although
 314 the LSTM based model demonstrates better performance on the training data from DIII-D
 315 plasmas, as shown in the left panel of Figure 4, the TCN based model achieves significantly
 316 better inference result for the JET data, as shown in the right panel of Figure 4. The optimal
 317 alarm threshold estimated base on the DIII-D validation data also generalizes much better to the
 318 JET test data for the TCN based model. This result indicates that the TCN based model has

319 learned deeper physics-based information, which is general for both plasma devices, and can
320 thus achieve disruption predictions for plasma conditions that it has never seen during training.

321
322 Figure 5 shows the signals from an example JET shot in the top four panels, and outputs from the
323 FRNN models trained on DIII-D database in the last panel, where only the TCN based model is
324 able to catch the continuously rising core radiation power as plotted in blue line in the second
325 panel, and predict for the imminent disruption, even when the input power is decreased, as
326 plotted in green line in the second panel. On the contrary, the LSTM based model output had a
327 sharp rise in disruption score at around 12s, when the input power is suddenly increased.

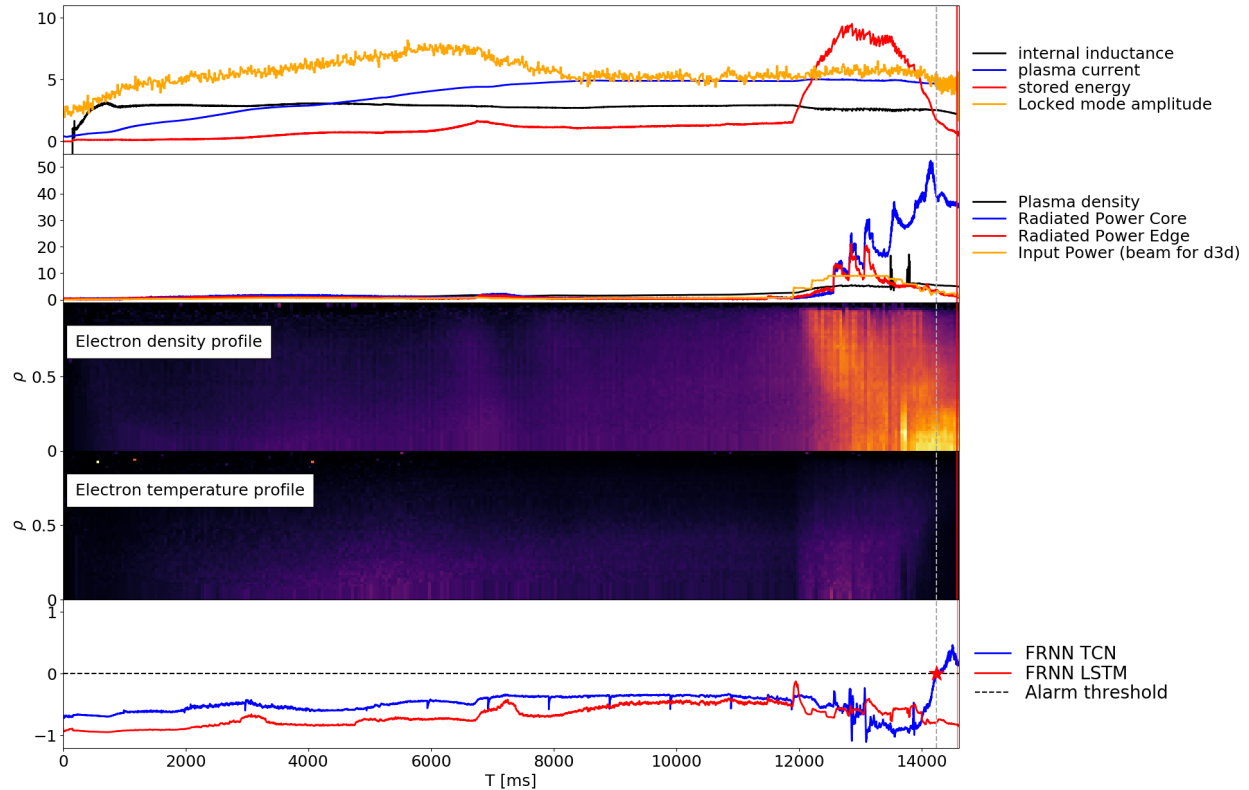
328
329 Although in the two aforementioned tasks, the TCN based models outperform the LSTM based
330 models in FRNN, we highlight here that the LSTM based modes can outperform the TCN based
331 models for other tasks, such as the prediction of disruptions in DIII-D plasmas when trained
332 based on JET plasma signals, as shown in the last column in Table 2. It is therefore important to
333 recognize that different types of deep learning architectures can be suitable for different types of
334 physics problems, and even different facets for the same problem, and the possible generally
335 improved predictive powers if the ensemble model based on multiple diverse deep learning
336 architectures are considered, especially for cross-machine predictive tasks.

337



338
339 Fig 4. Comparison of ROC curves on the DIII-D training (left panel) and JET test (right panel)
340 dataset for the optimal FRNN OD models, based on the TCN (blue) and LSTM (red)
341 architectures. The training and test score for the LSTM based model is 0.96 and 0.85
342 respectively. For the TCN based model, both training and test score is 0.91, indicating much less
343 overfitting. The solid dots indicate model performance at the optimal alarm threshold determined
344 on the validation set. The generalization of the optimal alarm threshold from DIII-D validation
345 set to JET test set is significantly better.

346



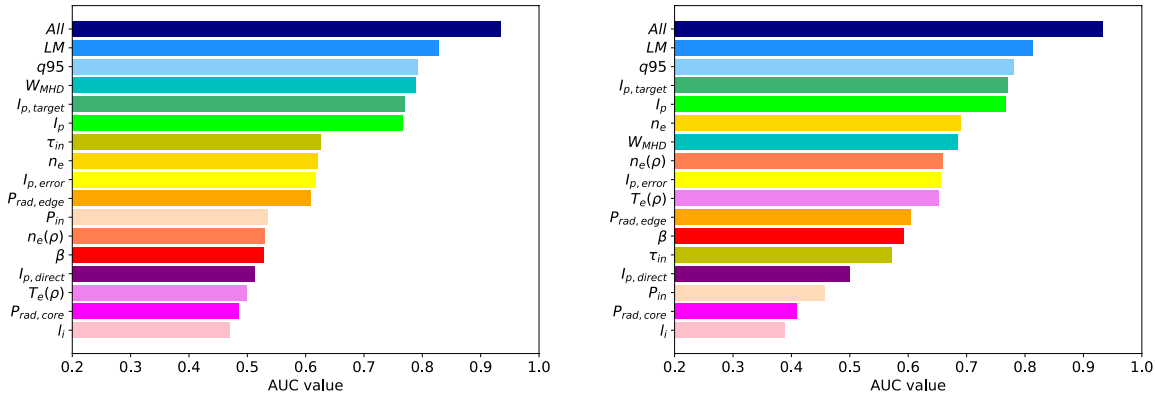
347
348

349 Fig 5. Example prediction on JET shot #83340. The top two panels show eight representative 0D
350 scalar signal channels (normalized by numerical scale on JET as listed in Table 1), and the next
351 two panels show two 1D profile signals channels as FRNN inputs. The last panel shows the
352 disruption scores returned by FRNN using TCN (blue line) and LSTM (red line) based models.
353 The signals are plotted up to the time of disruption. The dashed vertical line indicates the
354 disruption alarm time, and the solid vertical red line shows the latest warning time (30ms before
355 the disruption).

356

357 In order to achieve an improved level of understanding of the physics and signal features
358 underlying these disruption prediction results, we carried out a series of sensitivity studies.
359 Specifically, we assessed the signal importance of all input signals for the single machine
360 disruption prediction task on DIII-D using FRNN-1D. To assess the contribution from each of
361 the 16 signals to the disruption scores, we trained 16 individual models each using one single
362 signal at a time. In Figure 6, test performances are compared for the FRNN-1D TCN-based (left
363 panel) and the LSTM-based (right panel) models trained with the single labeled signal. As
364 expected, for both architectural schemes, the models trained with single channels have
365 significantly lower performance than the model trained with all 16 signals (represented by the
366 deep blue bar). The signal importance as measured by model performance for each of the two
367 models are affected by multiple factors, including initialization stochasticity and model
368 hyperparameters. Therefore, some variation in signal importance values between the models is
369 expected. However, it is important to note that there are clear qualitative trends that are
370 consistent in both panels of Figure 6. For example, the models trained on either the locked mode
371 amplitude (LM) or the tokamak safety factor value approaching the plasma periphery (q_{95})

372 signals outperform models trained on the rest of the signals, indicating that these two signals
 373 contain key disruption related information.



374
 375 Fig 6. Signal importance studies for models based on the TCN (left) and LSTM (right)
 376 architectures. Each bar represents the test set AUC values achieved by a model trained on the
 377 single labeled signal. The general trend of this sensitivity study is similar for the two
 378 architectures, showcasing the robustness of this method in estimating the relative importance of
 379 different physical signals.

380 In each panel in Figure 6, all models are trained using the same set of hyperparameters as those
 381 used for the respective best model tuned with all signals (represented by the deep blue bar), and
 382 34 models in total are trained for this analysis. In future investigations, more reliable estimates
 383 could be obtained by running hyperparameter tuning for each of these models, making it
 384 necessary to train thousands of models. Such a task would clearly require the associated
 385 engagement and effective utilization of modern supercomputers.

386 During real-time plasma experiments, due to unavailable diagnostic equipment, contaminated
 387 data collection process, or changing configuration of the experimental device and control system,
 388 some of the channels in the model inputs can be missing or containing high-amplitude noise.
 389 Therefore, we try to improve the robustness of our model performance with respect to input
 390 noise. There are many schemes to train noise-aware networks that can be stable against input
 391 noise (Seltzer and Wang, 2013; Zheng et al., 2016), here we use the simplest approach to
 392 perform noise-aware training by augmenting the training data with dropped out (set to zero)
 393 input channels to mimic unavailable experimental signals. The noise-aware models can
 394 significantly improve testing results when dropouts are added to the input data, at the cost of
 395 very small decrease in performance on the original full input. An example for DIII-D to JET
 396 cross machine predictive task is shown in table 4. When the channels in the input of the test
 397 dataset are dropped out with 0.1 probability (if the model is trained with 10 signals as input, then
 398 expectation of number of channels/signals that are set to zero is one) and 0.2 probability, the
 399 noise-aware model outperform the baseline model and is more robust on the perturbed data. The
 400 noise-aware training can help achieve predictive capabilities of the deep learning based models
 401 during real-time experiment in scenarios where certain diagnostics are unavailable or random
 402 data contamination happens.

403

Input channel dropout probability	0	0.1	0.2
Baseline model	0.902	0.847	0.798
Noise-aware model	0.900	0.875	0.839

405 **Table 4. Prediction results on JET dataset from a model trained on the DIII-D datasets, measured**
406 **as AUCs at 30ms before a disruption. Noise-aware models achieve significantly more robust**
407 **prediction results on inputs contaminated with dropped out channels. The models are trained on**
408 **11 input signals, among which 2 channels are 1-D profiles and 9 channels are 0-D scalar signals.**
409 **In the first column where input channel dropout probability is zero, the models are tested on the**
410 **original full input data. In the second and third column the input channels in the test database are**
411 **set to zero with probability 0.1 and 0.2. The noise-aware model is trained with corresponding**
412 **input dropout probability. In the first column the performance of the noise-aware model trained**
413 **with 0.1 input dropout probability is reported.**

414

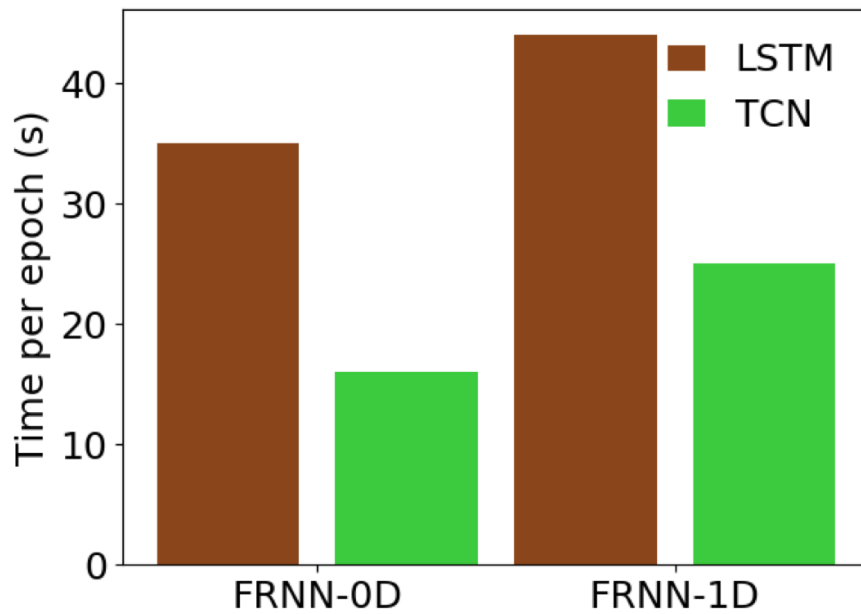
415 IV. Computational Performance Evaluation

416

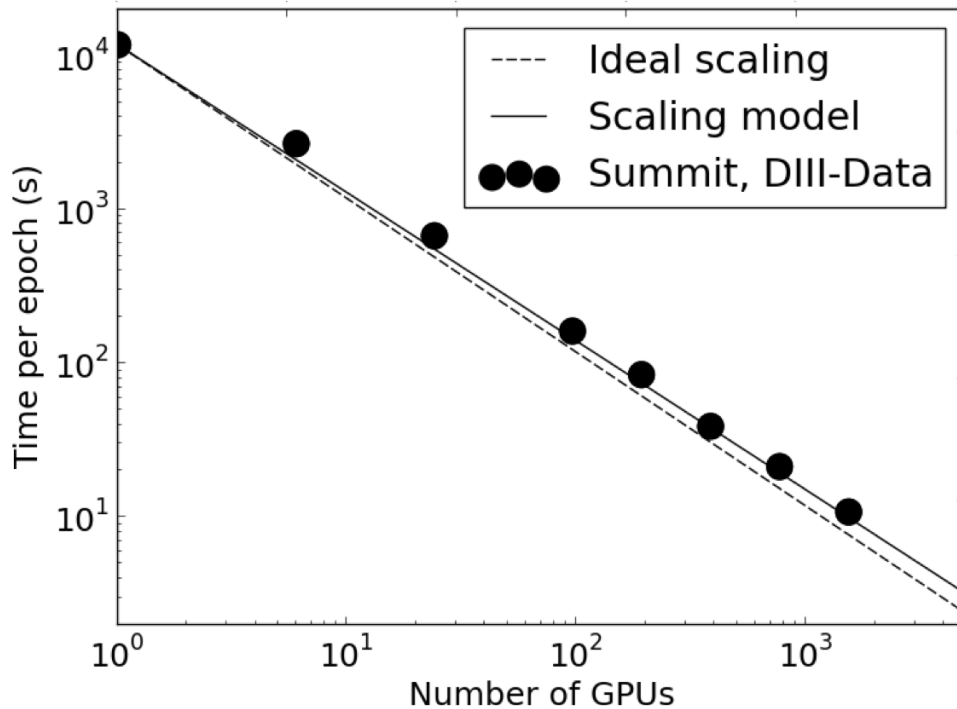
417 As demonstrated in this paper, the training and tuning of deep learning computational software
418 like FRNN requires modern supercomputing power that must be utilized with excellent
419 efficiency. Here, we demonstrate that with the new TCN architecture, FRNN exhibits even better
420 computational performance. Specifically, Figure 7 shows performance comparisons for the best
421 TCN-based and the best LSTM-based models, when both are trained on the DIII-D database with
422 30 ms warning time. For this task, the key finding is that the TCN architecture enabled FRNN to
423 reduce training time by about a factor of 2.

424

425 As shown in Figure 8, FRNN exhibits impressive strong scaling on the Oak Ridge Leadership
426 Computing Facility (OLCF) “Summit” – currently the top-rated supercomputer worldwide
427 (<https://www.top500.org/list/2019/06/>). Specifically, the training time for this TCN-based model
428 scales almost ideally with the number of GPUs used. In this scaling study, we used a much larger
429 database and a different set of model hyperparameters than that used in Figure 7, to avoid MPI
430 operation problems when the training time becomes too small for a large number of GPUs. This
431 important finding motivates the future development of modern deep learning software such as
432 FRNN in various aspects – including performance optimization via hyperparameter tuning,
433 consideration of more complicated and deeper architectures, and the addition of higher
434 dimensional input data sources with higher resolution.



435
 436 Fig 7. Time per epoch (i.e., the time required to complete one pass over the entire training
 437 dataset) during training using 4 Tesla V100s for FRNN-0D and FRNN-1D, for the LSTM
 438 (brown) and TCN (green) architectures, respectively. Lower values correspond to better
 439 computational performance. The four models studied here correspond to the best models from
 440 the studies of DIII-D single machine disruption prediction with 30ms warning time (see first
 441 column in Table 2).
 442



443 Fig 8. Strong scaling of FRNN with the new TCN-based model carried out on the Oak Ridge
 444 Leadership Computing Facility (OLCF) Summit system – currently top rated supercomputer
 446 worldwide. The time required to complete one epoch during training on Summit (black data
 447 points) agrees well with the scaling model. The original strong scaling of FRNN using the
 448 LSTM-based model and the associated derivation of the scaling model can be found in Kates-
 449 Harbeck et al., 2019.

450
 451 V. Summary and future work

452
 453 In this paper, we have introduced and implemented a new architectural scheme based on dilated
 454 temporal convolutions within the exemplar magnetic fusion plasma disruption prediction
 455 software FRNN. Compared with the previously published models based on the LSTM
 456 architecture, FRNN models constructed with the temporal convolutional neural network (TCN)
 457 architecture exhibit at least equivalent and demonstrably superior computational performance
 458 and predictive power for disruption forecasting across various experimental databases from the
 459 DIII-D and JET tokamaks. The TCN architecture has also been applied in other disruption
 460 studies recently, using input from the Electron Cyclotron Emission imaging (ECEi) diagnostic
 461 data on DIII-D (Churchill et al., 2020).

462
 463 In the present paper, we have developed a general deep learning capability to train multiple
 464 models with distinct architectures within a single software suite that serves to promote
 465 adaptability to different temporal and spatial learning tasks and also to enable ensemble schemes
 466 for highly accurate prediction. Various machine learning based algorithms targeting different
 467 learning and prediction tasks have been independently validated for modern tokamaks such as
 468 DIII-D (Rea et al., 2019) and JET (Rattà et al., 2010; Vega et al., 2014; Dormido-Canto et al.,

469 2013; Murari et al., 2013; Ferreira et al., 2019). The capability of utilizing effective ensemble
470 models in real-time plasma control systems is an important task for successful operation of future
471 machines. To contribute to this effort, we plan to implement additional deep learning based
472 architectures, including those based on attention-based models, such as the “Transformer”
473 (Vaswani et al., 2017), into the FRNN framework, as well as utilizing more high dimensional
474 experimental data. The goal is to develop FRNN into a flexible and adaptable software platform
475 for the prediction and analysis of complex plasma dynamics, including early disruption
476 prediction as well as other important physics phenomena. Beyond plasma device performance
477 predictions, this platform also has potential for wide applications to time series prediction
478 problems in other research areas, such as magnetosphere substorm onset predictions and weather
479 forecasts.

480 481 Acknowledgements

482
483 Research for this paper was carried out at the Princeton Plasma Physics Laboratory by the
484 Department of Energy (DOE) contract DE-AC02-09CH11466. The authors thank Professor
485 Zhihong Lin of UC Irvine for useful discussions and important support of this work associated
486 with the (DOE) SciDAC ISEP Center which he leads. We have benefited from the HPC
487 resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National
488 Laboratory (DOE Contract No. DE-AC05-00OR22725) and the National Energy Research
489 Scientific Computing Center (DOE Contract No. DE-AC02-05CH11231. This work is also
490 supported under Contract DE-AC02-06CH11357 associated with the Argonne Leadership
491 Computing Facility (ALCF) Aurora Early Science Program project at the Argonne National
492 Laboratory. The simulations presented in this article were performed partly on computational
493 resources featuring the “Traverse” cluster managed and supported by Princeton University’s
494 Research Computing Center, a consortium of groups including the Princeton Institute for
495 Computational Science and Engineering (PICSciE) and the Office of Information Technology
496 (OIT). We also express our gratitude to the EUROfusion Joint European Torus (JET) and their
497 management as well as to General Atomics (GA) and its DIII-D tokamak project for access to
498 the same fusion databases which were previously provided for the Nature (April, 2019)
499 publication. In addition, we extend special thanks to Dr. Nik Logan of PPPL for his careful
500 reading and associated helpful suggestions that improved the clarity of this manuscript. This
501 material is based upon work supported by the US DOE, Office of Science,
502 Office of Fusion Energy Sciences, using the DIII-D National Fusion Facility, a DOE Office of
503 Science user facility, under award DE-FC02-04ER54698.

504
505 Disclaimer: This report was prepared as an account of work sponsored by an agency of the
506 United States Government. Neither the United States Government nor any agency thereof, nor
507 any of their employees, makes any warranty, express or implied, or assumes any legal liability or
508 responsibility for the accuracy, completeness, or usefulness of any information, apparatus,
509 product, or process disclosed, or represents that its use would not infringe privately owned rights.
510 Reference herein to any specific commercial product, process, or service by trade name,
511 trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement,
512 recommendation, or favoring by the United States Government or any agency thereof. The views
513 and opinions of authors expressed herein do not necessarily state or reflect those of the United
514 States Government or any agency thereof.

515
516
517 References
518
519 Bai, S., Z. Kolter, J., and Koltun, W., An Empirical Evaluation of Generic Convolutional and
520 Recurrent Networks for Sequence Modeling, <https://arxiv.org/abs/1803.01271>, 2018

521 Bengio, Y., Simard, P., and Frasconi, P., Learning long-term dependencies with gradient descent
522 is difficult, *IEEE Transactions on Neural Networks* vol. 5, no. 2, 1994
523

524 Churchill, R.M., Tobias, B., Zhu, Y., and DIII-D team, Deep convolutional neural networks for
525 multi-scale time-series classification and application to tokamak disruption prediction using raw,
526 high temporal resolution diagnostic data, *Phys. Plasmas*, vol. 27, no. 062510, 2020

527 de Vries, P. C., Pautasso, G., Humphreys, D., Lehnen, M., Maruyama, S., Snipes, J. A.,
528 Vergara, A., and Zabeo, L., Requirements for triggering the ITER disruption mitigation system.
529 *Fus. Sci. Technol.*, vol.69, p. 471–484, 2019

530 Dormido-Canto, S., Vega, J., Ramírez, J. M., Murari, A., Moreno, R., López, J. M., Pereira, A.,
531 and JET-EFDA Contributors, Development of an efficient real-time disruption predictor from
532 scratch on JET and implications for ITER, *Nucl. Fusion*, vol. 53, no. 113001, pp. 8, 2013

533 Ferreira, D. R., Carvalho, P. J. and Fernandes, H., Deep Learning for Plasma Tomography and
534 Disruption Prediction from Bolometer Data, *IEEE Transactions on Plasma Science*, vol. 48, p.
535 36-45, 2019

536 Gerhardt, S. P., Darrow, D. S., Bell, R. E., LeBlanc, B. P., Menard, J. E., Mueller, D.,
537 Roquemore, A. L., Sabbagh, S. A., and Yuh, H., Detection of disruptions in the high- β spherical
538 torus NSTX, *Nucl. Fusion*, vol 53, no. 063021, 2013

539 Glasser, A. S., and Kolemen, E., A robust solution for the resistive MHD toroidal Δ' matrix in
540 near real-time, *Physics of Plasmas*, vol. 25, no. 082502, 2018
541

542 Hollmann, E. M. et al., Status of research toward the ITER disruption mitigation system, *Phys.*
543 *Plasmas*, vol. 22, no. 021802, 2015

544 Kates-Harbeck, J., Svyatkovskiy, A., and Tang, W., Predicting disruptive instabilities in
545 controlled fusion plasmas through deep learning, *Nature*, vol 568, pp. 526–31, 2019

546 Liu, D., Zhang, W., McClenaghan, J., Wang, J., and Lin, Z., Verification of gyrokinetic particle
547 simulation of current-driven instability in fusion plasmas. II. Resistive tearing mode, *Physics of*
548 *Plasmas*, vol. 21, no. 122520, 2014

549 López, J. M., Vega, J., Alves, D., Member, IEEE, Dormido-Canto, S., Murari, A., Ramírez, J.
550 M., Felton, R., Ruiz, M., de Arcas, G., and JET EFDA contributors, Implementation of the
551 disruption predictor APODIS in JET real time network using the MARTe framework, *2012 18th*
552 *IEEE-NPSS Real Time Conference*, Berkeley, CA, p. 1-4, 2012

553
554 Matthews, G. F. et al., JET ITER-like wall—overview and experimental programme, *Phys. Scr.*,
555 no. 014001, 2011

556 Murari, A. et al, Adaptive predictors based on probabilistic SVM for real time disruption
557 mitigation on JET, *Nucl. Fusion*, vol. 58, no. 5, 2018

558 Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N.,
559 Senior, A., and Kavukcuoglu, K., Wavenet, A generative model for raw audio,
560 arXiv:1609.03499, 2016

561 Pautasso, G. et al., On-line prediction and mitigation of disruptions in ASDEX Upgrade, *Nucl.*
562 *Fusion*, vol. 42, no. 100, 2002

563 Rattá, G.A., Vega, J., Murari, A., Vagliasindi, G., Johnson, M.F., de Vries, P.C., and JET EFDA
564 Contributors, An advanced disruption predictor for JET tested in a simulated real-time
565 environment, *Nucl. Fusion*, vol. 50, no. 2, 2010

566
567 Rea, C., Montes, K. J., Erickson, K. G., Granetz, R. S., and Tinguely, R. A., A real-time machine
568 learning-based disruption predictor in DIII-D, *Nucl. Fusion*, vol. 59, no. 096016, 2019

569
570 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat,
571 Deep learning and process understanding for data-driven Earth system science, *Nature*, vol 566,
572 p. 195–204, 2019

573
574 Schmidt, J., Marques, M. R. G., Botti, and S., Marques, M. A. L., Recent advances and
575 applications of machine learning in solid-state materials science, *npj Comput. Mater.*, vol 5, no.
576 83, 2019

577
578 Schuller, F., Disruptions in tokamaks. *Plasma Phys. Contr. Fusion*, vol. 37, no. A135, 1995

579 Seltzer, M. L., Yu, D., and Wang, Y., An investigation of deep neural networks for noise robust
580 speech recognition, *2013 IEEE International Conference on Acoustics, Speech and Signal*
581 *Processing*, Vancouver, BC, pp. 7398-7402 (2013)

582
583 Strait, E. J. et al., Progress in disruption prevention for ITER, *Nucl. Fusion*, vol. 59, no. 112012,
584 2019

585 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and
586 Polosukhin, I., Attention is all you need, *Advances in Neural Information Processing Systems*,
587 pp. 6000–6010, 2017

588 Vega, J., Murari, A., Dormido-Canto, S., Moreno, R., Pereira, A., Acero, A., and JET-EFDA
589 Contributors, Adaptive high learning rate probabilistic disruption predictors from scratch for the
590 next generation of tokamaks, *Nucl. Fusion*, vol. 54, no.123001, p. 17, 2014

591 Webb, S., Deep learning for biology, *Nature*, vol 554, p. 555-557, 2018

592
593 Zheng, S., Song, Y., Leung, and T., Goodfellow, I., Improving the Robustness of Deep Neural
594 Networks via Stability Training, *Proceedings of the IEEE Conference on Computer Vision and*
595 *Pattern Recognition (CVPR)*, pp. 4480-4488 (2016)
596
597 <http://www.ga.com/diii-d>
598
599 <https://www.euro-fusion.org/devices/jet/>
600
601 <https://www.top500.org/list/2019/06/>