PAPER

Alfvén eigenmode classification based on ECE diagnostics at DIII-D using deep recurrent neural networks

To cite this article: Azarakhsh Jalalvand *et al* 2022 *Nucl. Fusion* **62** 026007

IOP Publishing | International Atomic Energy Agency Nuclear Fusion

Nucl. Fusion **62** (2022) 026007 (11pp) https://doi.org/10.1088/1741-4326/ac3be7

Alfvén eigenmode classification based on ECE diagnostics at DIII-D using deep recurrent neural networks

Azarakhsh Jalalvand[1,2,*], Alan A. Kaptanoglu[3], Alvin V. Garcia[4], Andrew O. Nelson[5], Joseph Abbate[5,6], Max E. Austin[7], Geert Verdoolaege[8], Steven L. Brunton[9], William W. Heidbrink[4] and Egemen Kolemen[2,5,*]

[1] Department of Electronics and Information Systems, Ghent University, Ghent, B-9052, Belgium
[2] Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, United States of America
[3] Department of Physics, University of Washington, Seattle, WA 98195, United States of America
[4] Department of Physics, University of California, Irvine, CA 92697, United States of America
[5] Princeton Plasma Physics Laboratory, Princeton, NJ 08544, United States of America
[6] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, United States of America
[7] Institute for Fusion Studies, The University of Texas at Austin, Austin, TX 78712, United States of America
[8] Department of Applied Physics, Ghent University, Ghent, B-9000, Belgium
[9] Department of Mechanical Engineering, University of Washington, Seattle, WA 98195, United States of America

E-mail: Azarakhsh.Jalalvand@ugent.be and ekolemen@pppl.gov

Received 25 August 2021, revised 12 November 2021
Accepted for publication 22 November 2021
Published 17 December 2021

Abstract

Modern tokamaks have achieved significant fusion production, but further progress towards steady-state operation has been stymied by a host of kinetic and MHD instabilities. Control and identification of these instabilities is often complicated, warranting the application of data-driven methods to complement and improve physical understanding. In particular, Alfvén eigenmodes are a class of ubiquitous mixed kinetic and MHD instabilities that are important to identify and control because they can lead to loss of confinement and potential damage to the walls of a plasma device. In the present work, we use reservoir computing networks to classify Alfvén eigenmodes in a large labeled database of DIII-D discharges, covering a broad range of operational parameter space. Despite the large parameter space, we show excellent classification and prediction performance, with an average hit rate of 91% and false alarm ratio of 7%, indicating promise for future implementation with additional diagnostic data and consolidation into a real-time control strategy.

Keywords: DIII-D, electron cyclotron emission, Alfvén eigenmodes, reservoir computing networks, plasma control

(Some figures may appear in colour only in the online journal)

* Authors to whom any correspondence should be addressed.

1. Introduction

Most future magnetic-confinement-fusion reactor designs require steady-state operation for economic viability. In the context of high-performance tokamaks, steady-state operation necessitates active real-time control of a number of complex instabilities including edge-localized modes [1, 2], Alfvén eigenmodes (AEs) [3, 4], and more general disruptions [5, 6]. Even fundamentally steady-state devices such as stellarators will require sophisticated real-time control for modulating gas puff, divertor dynamics, and transport [7]. Furthermore, instabilities can occur on time scales of milliseconds or even microseconds. Subsequently, if plasma control schemes are built and updated in real-time, they are limited to simple models such as those based on 1D transport [8], linearization or local-expansions [9–18], heuristics (based on prior experimental knowledge) [19], the biorthogonal decomposition [20–24], and so forth. Many of these models have been successfully employed for real-time control in operational scenarios.

Experimental tokamaks such as DIII-D have decades of diagnostic data obtained in vast ranges of operational parameter space, so control models can be trained offline before consolidation with a real-time control algorithm. These large and expansive databases are optimal targets for data-driven approaches, which excel at extracting patterns from high-dimensional spaces [25]. Given that a high-quality database is available, machine learning techniques have distinct advantages over the aforementioned model types: (1) offline training allows for large nonlinear models, (2) multi-machine datasets facilitate finding universal plasma models across fusion devices [26], and (3) models can be generated for plasma instabilities and other plasma dynamics that are currently not well-understood or not amenable to any sort of linearization. Indeed, the variety and complexity of AEs in toroidal devices poses many challenges for simple models, generalization to new datasets, and analytic methods.

There has already been remarkable success in machine learning for disruption identification and real-time control in tokamaks [5, 6, 27–30], including high-performance models that are not limited to a specific device [26]. There has also been recent deep learning work for magnetohydrodynamic (MHD) and AE activity, which utilized manually-labeled spectrogram data from the TJ-II stellarator [31] and COMPASS tokamak [32] for automated identification of these modes in diagnostic data from a single magnetic probe. The former paper focuses on a binary classification of the spectrogram pixels, indicating whether each pixel corresponds to Alfvénic MHD activity or not. The latter focuses specifically on identifying a useful feature space for unstable reversed-shear Alfvén eigenmodes (RSAEs), which exhibit a unique frequency-sweeping behavior. A recent paper also showed that AE 'mode character' (i.e. whether the activity is chirping, avalanching, fixed frequency, or quiescent) can be effectively classified [33]. All three papers indicate promising avenues for future work. We improve on these initial papers in two ways: (1) the inputs in the initial studies are single spectrograms from magnetic probes, meaning there is no ability to use spatial correlations or identify internal modes that do not appear near the device walls, and (2) there was no attempt made at discrimination between different kinds of plasma dynamics.

1.1. Contributions of the present work

In contrast to the previous works, we utilize time-series from the 40-channel electron-cyclotron emission (ECE) diagnostic on the DIII-D tokamak to directly identify and classify AE activity from a set of five possible types indicated in table 1. This ECE diagnostic produces internal electron temperature measurements at 40 different radial locations, providing information about spatial correlations and capturing a wide range of internal modes. Our task is facilitated on DIII-D by a new labeled database of AE activity. We illustrate accurate AE classification and prediction performance with reservoir computing networks (RCNs), which are comparable or better than the current best performance rates in the field of machine-learning for plasma physics [6, 28, 34]. While we focus primarily on the identification of AE activity in the present work, in the future we expect to utilize the spatial and temporal information in the ECE diagnostics to extend our proposed models, determine the shapes and locations of AE modes in the plasma, and implement real-time control on DIII-D. The code used to produce this work is open-source at https://github.com/PlasmaControl and our AE database can be obtained by contacting the DIII-D team for data access.

1.2. Alfvén eigenmodes

AEs are a class of common instabilities observed in tokamaks and other plasma devices. Unfortunately, some types of AE instability, such as energetic particle resonance, can lead to confinement loss and damage to plasma-facing components of the device. The database used in this work (described in section 2) distinguishes between several types of AE activity: low-frequency modes (LFMs $\lesssim$ 50 kHz, these 'christmas light' patterns have been formerly characterized as BAAE modes [40]), beta-induced Alfvén eigenmodes (BAEs $\sim$ 30–150 kHz), ellipticity Alfvén eigenmodes (EAEs $\sim$ 150–200 kHz), reversed-shear Alfvén eigenmodes (RSAEs $\sim$ 100–200 kHz), and toroidal Alfvén eigenmodes (TAEs $\sim$ 90–200 kHz) [40, 48]. The quoted frequency ranges for each type are approximate, specific to DIII-D, and can vary significantly in differing DIII-D parameter regimes such as L-mode or H-mode. The AE modes are further described in table 1, where references to the relevant theoretical and experimental manuscripts can also be found. Energetic geodesic acoustic modes (EGAMs) [49] are also identified in this database but these modes typically require additional diagnostics such as magnetics to fully classify; since we focus on classification only via ECE, EGAMs are omitted in this manuscript.

Lastly, AEs are an excellent choice for training predictive models, because there are a wide range of experimental actuators that can be used for real-time control of different AE activity. Recent work indicates TAE suppression by resonant magnetic perturbations in the EAST tokamak [50] and AE stabilization in DIII-D via a controlled energetic ion density

Table 1. Description of the AE activity considered in this work, adapted from Heidbrink [47]. The poloidal wave number is denoted m and the minimum value of the safety factor is denoted $q_{\min}$.

Acronym	Name	Cause
BAE [35–37]	Beta	Compressibility
EAE [38, 39]	Ellipticity	m and $m+2$
LFM [40]	Low-frequency modes	Hot electrons, $q_{\min} \sim$ rational
RSAE [41, 42]	Reversed-shear	$q_{\min}$
TAE [43–46]	Toroidal	m and $m+1$

ramp [51]. For a review of potential AE control avenues, see Garcia-Munoz *et al* [52].

1.3. Electron cyclotron emission (ECE)

Electron cyclotron emission (ECE) provides direct local measurements of the electron temperature for thermal DIII-D plasmas [53], and as such, can provide spatiotemporally-localized information about AE activity. The electron temperature and all ECE data is reported in keV throughout the paper. The DIII-D ECE diagnostic data is obtained at 500 kHz, in a single toroidal cross-section, at $N_c = 40$ different channels corresponding to varying radial locations, as shown in figure 1. Each ECE channel spans an approximately 1–2 cm radial extent, which is small compared to AE structure (most of the time, an AE mode can be seen across several channels). For this preliminary work, we rely primarily on the strength of RCNs to capture temporal correlations for prediction. Analyzing the spatial correlations in the data (e.g. to find the location of instabilities) requires substantially more data pre-processing and this is likely a worthwhile change for follow up research.

Properly capturing the spatial correlations is difficult because the ECE radial positions change with the magnetic field, and therefore can vary substantially during startup operation. The first few ECE channels regularly view data that is outside the last closed flux surface (LCFS); this data is not a trustworthy measurement. In fact, any signal from outside the LCFS is not blackbody emission. At such locations, the measured emissions are typically a mix of downshifted X-mode radiation from the core, scrambled O-mode radiation, and other 'background' emissions. Although some plasma instabilities or features can sometimes be seen on these channels, the change in emissions means the measurements can no longer be interpreted as local. Despite this spatial variability and data corruption, this manuscript uses the full, raw, unprocessed ECE data, so that it uses only the ECE channel indices (i.e. only the relative radial positions of the measurements, not the absolute radial positions). This has the advantage that the magnetic field evolution is not required for our analysis.

A second potential complication in the data derives from the physics of ECE. In general, ECE data is not well-defined if the ECE measurement frequency is below the plasma cutoff given by [54],

$$\omega_R = \frac{\omega_{ce}}{2}\sqrt{1+4\frac{\omega_{pe}^2}{\omega_{ce}^2}}, \qquad (1)$$

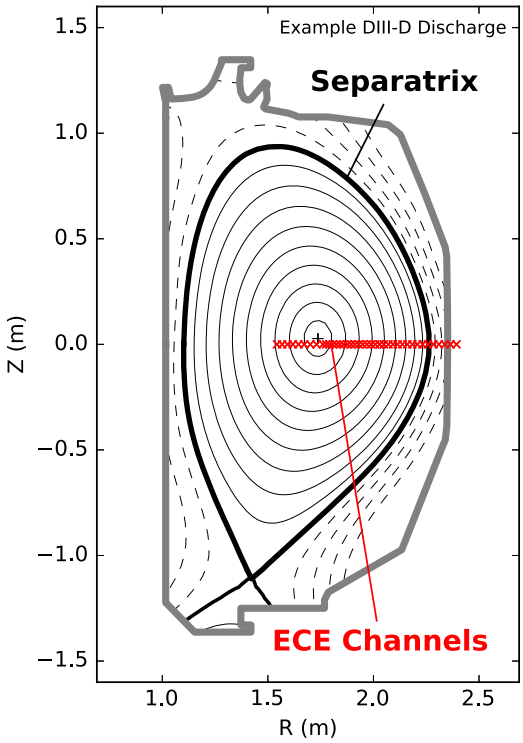

Figure 1. Illustration of the 40 radial ECE measurement locations alongside the closed (solid) and open (dashed) flux surfaces for an example DIII-D discharge. The ECE radial locations can vary significantly in each discharge, and measurements outside the LCFS are not local or accurate.

where ω_{ce} and ω_{pe} are the electron gyrofrequency and plasma frequency. Computing ω_R then requires external knowledge or measurements of the magnetic field and density profiles. Fortunately, these profiles can be estimated on DIII-D in real time so that the cutoff can be quickly evaluated. To evaluate the importance of these cutoffs, every 50 ms we have computed ω_R for each ECE channel and discharge. In our labeled database, we consider only the relatively low-density period before 1.9 s, corresponding to the times that are labeled for training. With these choices, the total number of cutoffs (summed over all the discharges, time slices, and ECE channels) that occur is approximately 500, meaning that the cutoff rate is a mere 0.03%. For this initial work, we have not discarded these very rare occurrences of ill-defined data. The error introduced from cutoffs is negligible compared to the label uncertainty, described in the following section.

Table 2. Mode characterization flags used to label the AEs in the dataset. During the training, we only consider the clear instances of AE modes which are marked in this table. All the other flags have been treated as *no AE*.

Flag	AE				
	BAE	EAE	LFM	RSAE	TAE
(3) AE with chirping	✓	—	✓	—	—
(2) AE without chirping	✓	✓	—	✓	✓
(1) marginal AE	—	—	—	—	—
(0) No AE	—	—	—	—	—
(−1) undetermined	—	—	—	—	—

In summary, we train machine learning models directly on the full, raw, unprocessed ECE time series data, including rare measurements below the cutoff and corrupted measurements from outside the LCFS. Additionally, we do not track the magnetic field evolution or the time evolution of the exact radial locations of each channel. In other words, we do not remove corrupted measurements nor do we penalize such instances during training. Despite these simplifications, we illustrate high classification and prediction performance in section 3.5, and it is interesting that high performance is accessible with minimal data processing.

2. The 2009–2017 DIII-D AE energetic particle database

A high-resolution, informative, and properly-labeled database is typically required for effectively training advanced machine learning models. In this regard, we have developed a current ramp AE dataset based on Heidbrink *et al* [40]. The labels in this DIII-D database cover many years of operation and a very broad parameter space. The database contains labels at fixed time snapshots indicating subjective flags for the corresponding AEs. The flags score the data as *clear AE (with or without chirping)*, *marginal AE*, *no AE*, or *undetermined*, meaning that there was not enough data to confidently classify the mode. To ensure that the classifier is trained only with strong evidence of AE, we only consider *AE without chirping* as examples of AE modes. The remaining flags are ignored and their corresponding datapoints are simply treated as *no AE*. However, we make two exceptions to this transformation for BAE and LFM since historically these modes are somewhat more challenging to identify. *AE with chirping* is also considered for BAE, and only this flag is taken into account for LFM (see table 2). An example of converting these flags to training target is also presented in figure 5.

The dataset consists of 1139 discharges collected between years 2009–2017, although in this work we focus on a subset of random 600 discharges because it was empirically found that adding more discharges only marginally improves performance for this specific task. Figure 2 depicts the database AE labels for the DIII-D discharge 170670 superimposed on several spectrograms of the more illustrative ECE signals. We also add special marks to indicate the various AE modes so that the reader can visually identify the different AE mode

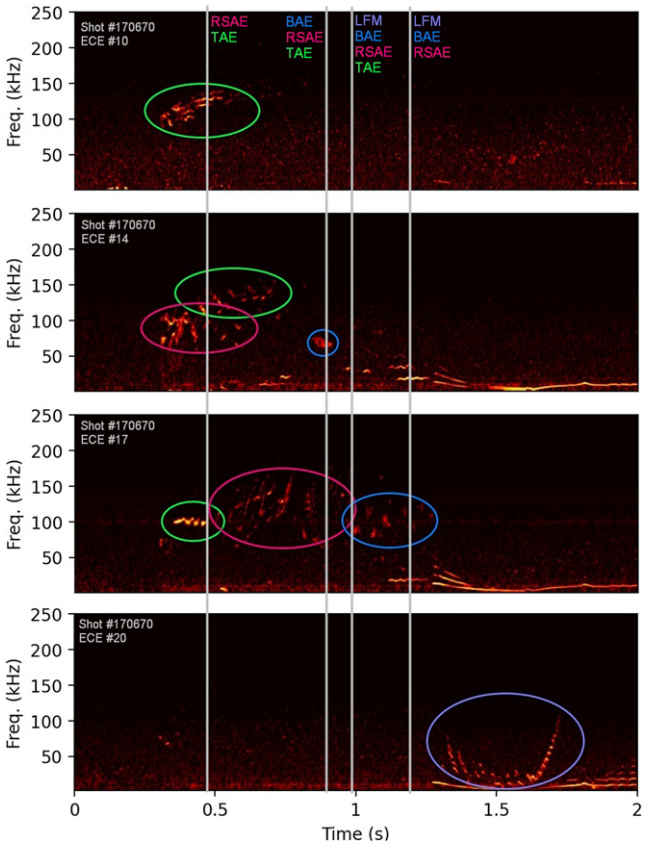

Figure 2. Illustration of several post-processed (denoised) ECE spectrograms for discharge 170670. The vertical white lines and labels indicate the database timestamps and corresponding instabilities that are used for training the model. The labels indicate only approximate occurrence and there can be substantial regions of unlabeled AE activity. For the reader, we also added some extra colored circles to this image to better visualize the different plasma modes.

types. In order to correctly label AE activity in the database, we often used a few different experimental diagnostics to cross-validate our label choices, especially when concurrent AEs were present. To ensure a variety of q profiles and to facilitate mode classification, selected times in the discharge are all during the first 1.9 s of the discharge, when the q profile steadily evolves [40]. Selected shots had a wide variety of purposes but nearly all dedicated energetic particle experiments are included. Time slices for the labels are chosen to sample either different plasma conditions or different types of mode activity, thus, a given discharge may have only a single AE label or as many as nine labels. In total, the database spans conditions including plasma current $I_p \leqslant 1.6$ MA, toroidal field $0.5 \leqslant B_T \leqslant 2.1$, normalized beta $0.1 \leqslant \beta_N \leqslant 3.2$, elongation $1.1 \leqslant \kappa \leqslant 2.2$, triangularity $-0.4 \leqslant \delta \leqslant 1.0$, line-average density $0.4 \times 10^{19} \leqslant \overline{n}_e \leqslant 5.0 \times 10^{19}$ m^{-3}, central electron temperature $T_e \leqslant 7.6$ keV, and central ion temperature $T_i \leqslant 11.4$ keV. Plasmas in both L-mode and H-mode are included. All discharges utilize deuterium neutral beam injection into a deuterium plasma and carbon is the dominant impurity in the graphite-wall vessel. More details about the labeling process can be found in [40].

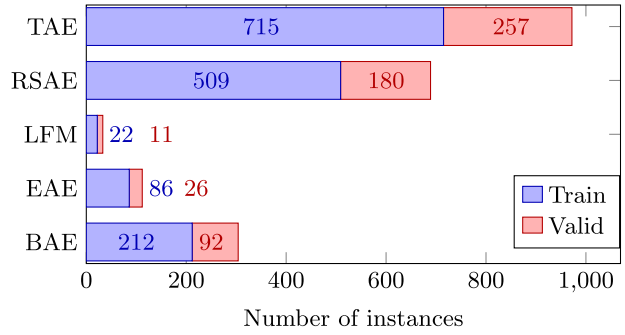

Figure 3. Number of instances for each of the manually-labeled AE types, for a random assignment of 450 training and 150 validation discharges. The distribution is heavily skewed towards RSAE and TAE activity.

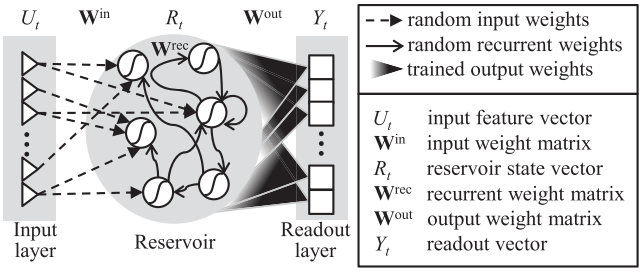

Figure 4. A basic RCN is composed of interconnected non-linear neurons with randomly fixed weights. The readout layer consists of linear neurons with trained weights.

As can be seen in figure 2, a main challenge in the available dataset is that the labels represent only a rough timestamp for each encountered event, indicating our confidence that AE activity has occurred somewhere in this vicinity. This means that the start, duration, and end of the AE activity is not clear from the labels. Therefore, we must define an arbitrary window around each AE timestamp and *assume* that all the datapoints in this window belong to that event. Such sub-optimal labels hinders perfect training and evaluation of the model. We visually inspected several discharges to estimate the typical temporal intervals of each of the AE modes. These approximate intervals vary between 50 to 500 ms, depending on the AE mode. For example, RSAEs and TAEs usually last longer than EAE, LFM and BAE. Therefore, we empirically defined a window of $\pm$125 ms around each labeled AE to create the targets for training the model.

For evaluation, we define custom performance metrics in section 3 that consider the model successful if it predicts AE activity within a determined window around the provided label. If the model triggers in more than 10% of this interval, it is considered as a true positive (TP).

The last note about the dataset is that the amount of AE activity of each type varies dramatically, as can be seen in figure 3. Also, LFMs are only labeled in the chronologically recent shots, so a random shuffle of the discharges is required before choosing training and validation sets.

In the next section we motivate and describe our AE classification model architecture based on this dataset.

3. AE classifier model and performance

Recurrent neural networks (RNNs) excel for the temporal analysis of a set of signals [55] and RCNs [56], are derived from more general RNNs. A simple RCN is a neural network with three particular computational layers: (1) the input layer, (2) a pool or 'reservoir' of non-linear neurons, driven by inputs and by delayed feed-backs of its outputs and (3) a 'readout' layer of linear neurons, driven by the hidden neuron outputs (figure 4). A fundamental point is that the input weights and the recurrent connection weights are initialized by random values, and that only the output weights are optimized (trained) using regularized linear regression for solving the targeted problem.

The intuition for this method is that the randomly initialized reservoir projects the input feature space to an untrained but typically much larger non-linear feature space in which the data samples can often be more accurately separated. Although typical RCNs are significantly less complex than state-of-the-art deep learning models, such as convolution neural networks, RCNs have shown comparable performance to these models for complex classification and prediction tasks [57, 58]. More details about the RCN parameters and properties can be found in appendix A and Jalalvand *et al* [57].

3.1. RCN-based model for AE classification

An overview of the data processing for AE classification is illustrated in figure 5. As we described before, normalizing the 40 ECE signals to have zero mean and unit variance is the only pre-processing step between the raw data and the input to RCN. In this step we also convert the table of AE labels to the targets required for training the model. Therefore the input to the RCN model at each time step t is U_t, a vector of size 40×1 and the target at that time is D_t, a vector of size 5×1 indicating which AE modes are present at that time.

It has also been shown that stacking RCNs improved the results for audio and image recognition because subsequent RCNs are able to learn and modify the errors provided by the previous ones [58, 59]. For this reason, we use a stack of two RCNs with the primary purpose of the second layer to smooth and better discriminate the final outputs.

We report the performance of our classifiers using the typical confusion matrix measures [60] per AE mode and/or over all AE modes together. Recall that the time labels provided for the dataset do not include the start and end boundaries of each AE event and also the time label could be anywhere during the occurrence of an event. This limitation motivates the following custom evaluation metrics. The detected AE times were compared to the reference labels. For each of the five AE modes, if the detected AE matches the manually-labeled AE in a time-window of $\pm$250 ms around the labeled AE, it was considered as one TP for the whole window, otherwise, a false negative (FN). This 500 ms window is twice as long as the window we considered around each label for training the RCN (see section 2). The manual labels approximately indicate when we were most confident about an event, so we would like to train the model with the input data as close as possible to the provided time-label. However, during the evaluation it

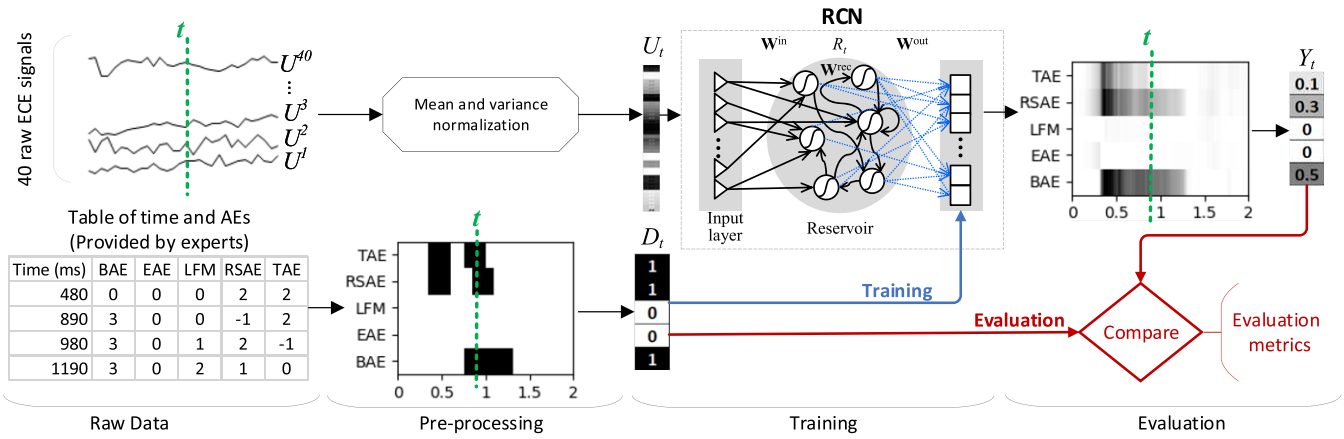

Figure 5. Diagram of the RCN-based data processing pipeline for AE classification. At each time step t during the training, the RCN is supplied with U_t, the normalized ECE signals of size 40×1 and the target at that time is D_t, a vector of size 5×1. The integer numbers in the 'raw data' table refer to the flags in table 2. After training the RCN output weights (dashed blue arrows), the output of the RCN is compared with the actual targets to evaluate the performance by measuring the TP and false positive (FP) rates.

would be fair to give the model some freedom to look for the AE modes in a wider window around each provided label and to be rewarded if it raises the correct flag. Outside this window, the readouts are expected to remain neutral. Therefore, if any AE was detected at each time step outside the window, it is considered as a false positive (FP). With these modified definitions of TP and FP, the measurements true positive rate (TPR) and false positive rate (FPR) are defined as usual through,

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{2}$$

In order to maximize our model performance, we investigate how performance changes with varying (1) input normalizations, (2) data sampling rates, and (3) model architecture, before summarizing our best model performance in section 3.5.

3.2. Inputs for the RCN

We begin with optimizing a rather small reservoir of 500 nodes and study different possible pre-processing of the 40 ECE input features. Figure 6 plots the TPR against the FPR over all AEs at various threshold settings (aka ROC curves) when the model is supplied with (1) *raw* ECE data, (2) *globnorm*: ECE data normalized over all channels in all training shots, (3) *globperchann*: ECE data normalized per channel over all training shots, and (4) *pershotperchann*: ECE data normalized per channel per shot. The best possible model would yield a point in the upper left corner or coordinate (0, 1) of the ROC space. A random guess would give a point along a diagonal line. This experiment shows that, in general, normalizing the data improves the performance of the model, which is typical for neural networks. Moreover, global normalization has similar performance to normalizing ECEs per channel per shot. It is useful that global normalization performs well. In a real-world scenario (while a discharge is running), we do not have an overview of the complete shot, so that per shot normalization is not feasible. Based on this experiment, global normalization

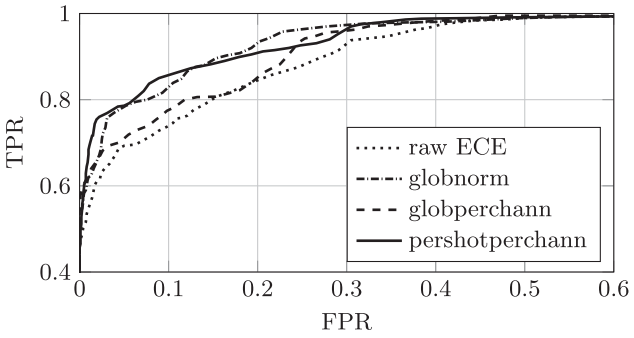

Figure 6. ROC curve when an RCN with 500 nodes is supplied with (1) *raw* ECEs, (2) *globnorm*: ECEs normalized over all channels and all training shots, (3) *globperchann*: ECEs normalized per channel over all training shot, and (4) *pershotperchann*: ECEs normalized per channel per shot.

is enough for the reservoir to capture the most useful information from the input features. For the remainder of this work, we globally normalize the ECE data by calculating one pair of mean and standard deviation over all ECE channels in all training discharges, and then use this pair to feature-scale both the training and validation data. There is no additional feature engineering on the raw ECE signals. We also do not incorporate any physics-based knowledge process, such as determining cutoff frequencies or invalid diagnostics, to remove or modify the dataset.

3.3. Impact of ECE signal resolution

In section 1.2 we discussed that traces of instabilities such as AEs are spread over a wide frequency range up to 200 kHz. Such complex patterns are only visible for human eyes on spectrograms, whereas it is almost impossible for human experts to detect them from the raw ECE time-series. This justifies the recording of ECE signals at the frequency of 500 kHz. On the other hand, our experiments showed promising performance of the proposed model in learning these patterns from the raw signals. Consequently, it is interesting to

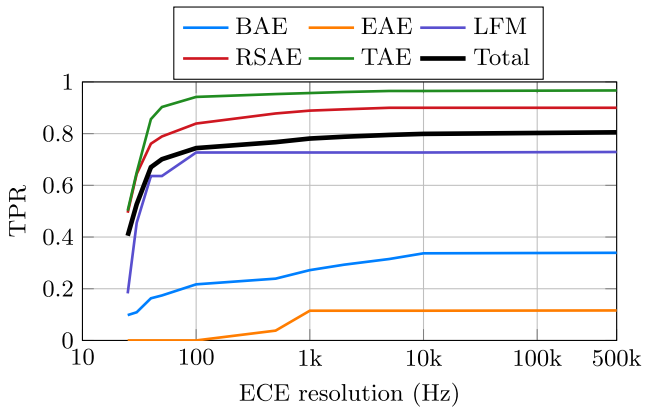

Figure 7. TPR per AE as a function of the input ECE resolution for a single-layer RCN with 500 nodes. Detection of the shorter AEs such as BAE and EAE is more dependent on the high-resolution ECE inputs compared to AE modes with a longer duration.

study the impact of the data resolution on the performance of the model. In that regard, we used the systematic random sampling approach [61] and gradually reduced the resolution of ECE signals by keeping only one datapoint out of each window of n datapoints. For example, $n = 1000$ means that only one out of every 1000 datapoints of each ECE signal has been kept, hence, the original sampling rate of 500 kHz is reduced to 500 Hz.

Figure 7 plots the TPR for each AE class as a function of the ECE resolution (measured by the sampling rate) for an RCN with 500 reservoir nodes. It was surprising to observe that reducing the sampling rate from 500 kHz down to 1 kHz does not produce a significant reduction in AE detection. It is only for sampling rates below 1 kHz that the performance gradually drops. In these experiments, both training and validation data were subsampled to the same sampling rate. However, we observed that for example, training a model at 2 kHz and evaluating on 1 kHz up to 500 kHz data did not show notable performance degradation.

Some hypotheses for this behaviour are as follows:

- In our dataset, the two most frequent instabilities, RSAE and TAE typically last for a few hundred milliseconds. Hence, even at a 1 kHz sampling rate there are several hundred datapoints representing the AE activity. On the other hand, less frequent instabilities such as BAE and EAE are usually shorter in time; in these cases, the detection of such instabilities are more sensitive to the correct sampling rate. Also unstable EAEs usually occur at frequencies that are outside the ECE bandwidth, hence, the model has difficulties in detecting them. The results in figure 7 are inline with this hypothesis.
- Machine learning models such as RNN and RCN which benefit from a so-called *memory* are usually more tolerant against missing information in time. However, our follow-up experiments showed that although discarding the memory slightly decreased the performance, the memory-less model still follows the same trend of behavior against changing the sampling rate. This typically means that the model is primarily using the recent or immediate signal values rather than the full history of the waveform to reach decisions.
- From the hardware point of view, the 40 ECE probes are installed close to one another and their radial locations can vary significantly in each discharge. Therefore, it is likely that the collected information by these probes overlap both in the spatial and temporal domain [62]. As a result, the missing information in one probe might be covered by the neighboring probes.

Considering these experiments, the results in section 3.5 were obtained from ECE data subsampled to 2 kHz. This subsampling minimally impacts the classification performance, while reducing data processing expenses drastically. The preliminary results in this work show promising potential in detecting AE instability using low resolution diagnostics. Such a large data reduction enables the implementation of low-power, compact and real-time control modules on hardware with limited data processing capabilities, such as field programmable gate arrays. But undoubtedly, deeper investigation is required to draw concrete conclusion on the performance and reliability of this approach.

3.4. RCN size and depth

We aim at developing a multi-layer RCN that is obtained by stacking multiple RCNs [63]. The first RCN is supplied with the ECE data and trained with the provided labels. After training the first layer, the second reservoir is fed with the first readouts and is trained with the same labels used for training the first layer. By stacking reservoirs, the temporal modeling capacity of a single layer model is extended. In [58, 59], it was shown that this improved the results for audio and image recognition, and for multipitch tracking, because subsequent layers are able to learn and modify the errors provided by the previous layers.

Figure 8 shows TPR and FPR as a function of reservoir size and depth. We see that increasing the size of a single RCN from 50 to 8000 significantly boosts the TPR of the smaller classes such as EAE and BAE. The FPR is relatively less influenced by the size of reservoir but is fairly low already. Moreover, adding the second RCN, and increasing its size from 50 to 1000, considerably improves the FPRs of TAE and RSAE with minimal change to the TPRs. This is in line with previous conclusions that the major impact of increasing the size and depth of RCNs is to fix the misclassifications and false alarms without significantly altering the correct decisions [64].

The results suggest that a two-layer RCN with 8000 and 500 neurons for the first and second layer is sufficient. Such a model consists of only $5 \times (8000 + 1) + 5 \times (500 + 1) = 42,510$ trainable parameters. With a set of 1.7 million training data points, training of the first layer completes in 126 min, while training the second layer required only 30 min on a single IBM POWER9 CPU core.

3.5. Final results

Figure 9 plots TPR and FPR as a function of the threshold to detect each of the five AE modes. Since TAE and RSAE are the largest classes, it is expected that the classifier learns them

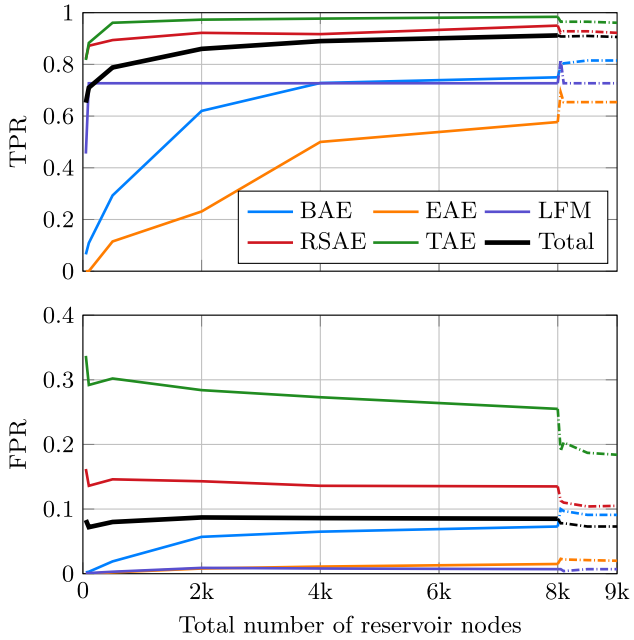

Figure 8. RCN performance as a function of reservoir size and depth. The solid lines show the TPR and FPR over all AEs when the single-layer reservoir size is increased from 50 to 8000. The dashed lines show the same measures when a second reservoir is added and trained on the output of the single-layer 8000-node model and its size increases from 50 to 1000. The second layer clearly improves the FPR without significantly affecting TPR.

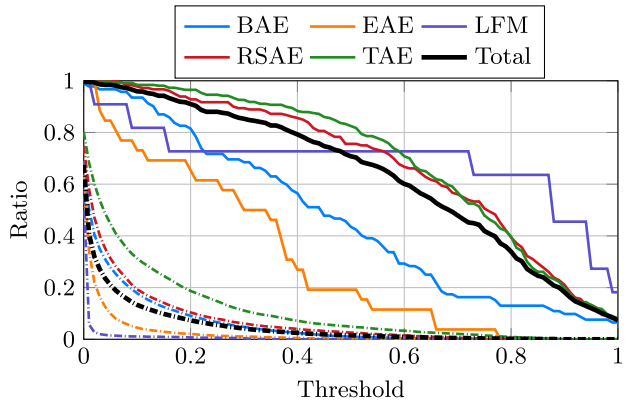

Figure 9. TPR (solid) and FPR (dashed) per AE of a two-layer RCN with 8K-500 nodes per layer on the validation set.

better than the others and they also influence the average TPR the most. BAE, EAE and LFM have lower hit rate, and among them EAE seems to be most challenging one. One possible explanation for the poor model performance on EAEs is that, in our DIII-D database, EAEs are significantly more temporally abrupt than the other AE modes. At a prediction threshold of 0.2, we obtain overall TPR = 0.91, FPR = 0.07. These rates are comparable or better in performance to the best rates in disruption prediction [6, 28, 34].

Table 3 lists the performance of the two-layer RCN with 8K-500 nodes per layer on the validation set. The table shows the imbalanced nature of the problem; in total, only 0.02% of the validation set are labeled as an AE mode (the training set

Table 3. Performance of a two-layer RCN with 8K-500 nodes per layer on the validation set. A threshold of 0.2 has been applied to binarize the model output. There are 566 labeled AEs in total.

AE	TP	FP	TN	FN	TPR	FPR
BAE	75	46 982	470 368	17	0.82	0.09
EAE	17	11 976	566 324	9	0.65	0.02
LFM	8	4102	587 088	3	0.73	0.01
RSAE	167	48 319	417 211	13	0.93	0.10
TAE	248	76 133	330 057	9	0.97	0.19
Total	515	187 512	2371 048	51	0.91	0.07

exhibits a similar percentage). This fact is why TPR and FPR are our primary metrics. Reporting the accuracy of the model would be profoundly misleading; a model that *never* predicts AE activity would report accuracy above 99%.

Moreover, there appears to be a clear correlation between the number of each AE mode and the corresponding model performance. TAE, RSAE and BAE have the highest TPRs. However, the higher TPR modes, such as TAE, also have slightly higher FPR. Two straightforward explanations for these results are poor training data and 'over-learning' in the model. However, there are additional two contributions to the higher FPR, which are difficult to disambiguate: (1) the model is identifying existing AE modes that are not labeled in the dataset and (2) some AE periods are longer than the arbitrary 500 ms window that we considered around each label. Consequently, the model can correctly raise a flag on the continuation of these events beyond the prescribed window, but the evaluation measure incorrectly reports a FP. We will discuss an illustrative example of this behavior later in this section. But more accurate labeling and evaluation process as well as further improvement for more balanced performance across the different AE modes are certainly avenues to move forward in the future work.

For comparison, we also trained alternative machine learning models including multi-layer perceptron, RNN, GRU and LSTM-based models, using the same dataset. None of these models could outperform the RCN; the best result was achieved by a three-layer LSTM with 256, 128 and 64 nodes (551000 trainable parameters in total) leading to TPR and FPR of 0.63 and 0.39, respectively. One possible explanation for this poor performance is that complex models have the capacity to learn many details from the available data and are subsequently more sensitive to the quality and quantity of the training data. Simpler models like the RCN are less sensitive. The current dataset is an illustrative example of this tradeoff, because AEs appear in only a small fraction of the data, the AE classes are highly imbalanced, and the provided labels are not ideal for training predictive models. Nevertheless, the strong performance with RCNs is encouraging for continued machine learning work in the future.

To provide an idea about the distribution of the AE events and detection, figure 10 presents the hit and miss on the validation dataset sorted chronologically. It shows that only a few LFMs have been labeled on the most recent shots. Another interesting observation is that most of the missing

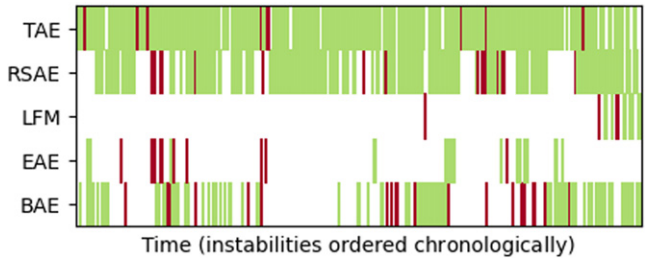

Figure 10. Hit (green) and miss (red) of detecting AE modes on the validation dataset sorted chronologically.

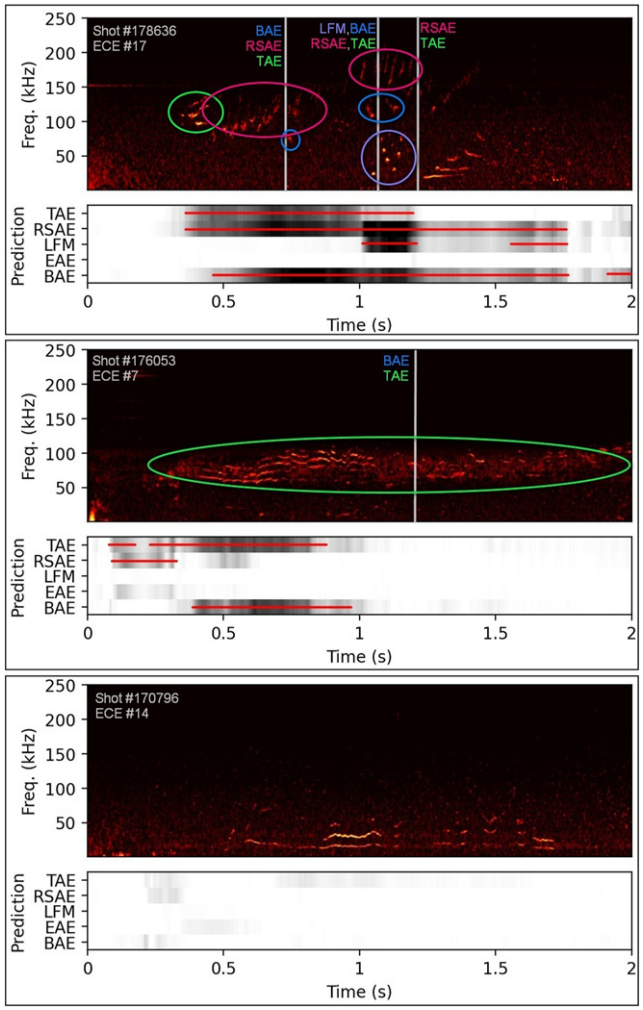

Figure 11. RCN outputs for discharges 178636 (top), 176053 (middle), and 170796 (bottom). For each discharge, the outputs of the final RCN model along with the denoised spectrogram of a selected ECE channel is plotted. The spectrograms are enriched with the labeled AE and time stamp in white, as well as extra color circles for better visualization of the AE patterns. Note that the spectrograms are for a single ECE channel while the output predictions are based on all 40 ECE channels as input to the model—some mismatch is expected. Outputs greater than a threshold of 0.2 are marked by red lines, indicating that the model has raised a positive flag for the AE. In discharge 176053, the proposed model nicely captures some unlabeled activity in the beginning of the discharge, but clearly misses some of the AE activity later in time. For discharge 170796, which has not been labeled with any evident AE mode, the model output correctly remains neutral.

EAEs occurred in the early shots, although this may just be a statistical fluke given too few EAE samples.

Finally, figure 11 illustrates two examples of the validation set, along with the corresponding labels and the model outputs. Shot 178636 is a complex example with several concurrent and labeled AEs. The RCN correctly predicts all of the labeled AEs. On the other hand, Shot 176053 is an example in which the model decision is very far from the provided BAE and TAE label at time 1250 ms. Interestingly, the model clearly suggests TAE and BAE between 400 ms and 1000 ms, which can be readily confirmed in the spectrogram. Moreover, revisiting the diagnostics confirmed that there is also evidence of RSAE at the beginning of this shot. Despite the disagreement with the labels, this example may paradoxically suggest strong model performance. The model 'errors' can be visually confirmed to be often an improvement over the true labels, illustrating that the model is correctly learning the AE features. An illustrative exception is the continuation of TAE in the second half of shot 176053, which the proposed model fails to identify. While it is generally difficult to interpret the behavior of black box models such as neural networks, it is worth noting that the performance of a conventional data-driven model strongly depends on the quality and quantity of the available training data.

Finally, we examined the behaviour of our proposed model on discharge #170696 which has not been labeled for any evident AE mode. While there are some non-AE activities during this shot, the RCN model confidently remains below the threshold of 0.2, meaning that the model is quite robust against the activities which it has not been trained for.

4. Conclusion

We have illustrated that simple and effective machine learning methods, such as RCNs, can excel at the classification and prediction of AEs directly from raw DIII-D ECE data. The available labels only roughly determine the timestamp in which an AE mode has occurred. Nevertheless the RCN-based model showed a promising hit rate of 0.91 in detecting five different AEs, and visual inspection of the prediction indicates that the model is correctly learning the features of AE activity. Moreover, our experiments suggest that subsampling the ECE signals from 500 kHz to as low as 2 kHz does not have significant influence on the performance of the classifier. This is of great importance for enabling real-time data reduction and featurization at the source of such high frequency, high bandwidth diagnostic signals.

Although this research provides a good proof of principle test showing the capability of simple yet effective models in identifying AE modes based on ECE diagnostics, there remains a lot more work to develop and improve such processing pipeline to a form usable in fusion reactors. For instance mapping the channels to radial location to correctly capture the spatial dependence in the ECE data can help to improve the AE classification performance or even to detect the location of the instability inside the plasma. Enriching the input data by a suite of diagnostics including ECE, BES, and magnetics, as well as investigating more complex deep learning methods are also other paths to more robust detection of the

instabilities. Future work with multi-machine datasets should also investigate building universal AE detection models for application across toroidal plasma devices. Lastly, preliminary experiments with spatially-localized convolutional neural networks indicate promising performance which we plan to further investigate and hopefully report in the future.

Disclaimer

Acknowledgments

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 Award(s) DE-FC02-04ER54698, DE-SC0021275, DE-SC0020337, Army Research Office (ARO W911NF-19-1-0045), National Science Foundation under 1633631 and Ghent University Special Research Award No. BOF19/PDO/134.

Appendix A. RCN equations and hyperparameters

RCNs are effective and robust for capturing temporal information in multi-variate time series data [58, 65, 66]. A basic RCN consists of the following components:

- The input weight matrix $\mathbf{W}^{\text{in}}$ passes the N^{in} features to the reservoir with N^{res} neurons.
- The reservoir weight matrix $\mathbf{W}^{\text{rec}}$ interconnects the neurons inside the reservoir.
- The output weight $\mathbf{W}^{\text{out}}$ connects the neurons to the output node.

Typically, $\mathbf{W}^{\text{in}}$ and $\mathbf{W}^{\text{rec}}$ are initialized from a random distribution between ± 1. Furthermore, they are typically initialized sparsely, which means that each neuron inside the reservoir receives a very small fraction of $K^{\text{in}} = 10$ input features and previous outputs of $K^{\text{rec}} = 10$ reservoir neurons. Conventionally, $\mathbf{W}^{\text{rec}}$ needs to fulfill the *echo state property* [56]. This ensures that the reservoir forgets about the past eventually and controls the impact of the memory on the neurons current activation. Therefore, $\mathbf{W}^{\text{rec}}$ is normalized to its maximum absolute eigenvalue and re-scaled by the hyper-parameter $\alpha_R < 1$ (spectral radius). The hyper-parameters α_U (input scaling) and α_R together control the balance between the impact of the new inputs and the memory on the reservoir outputs.

The key difference between RCNs and typical RNN architectures is that $\mathbf{W}^{\text{in}}$ and $\mathbf{W}^{\text{rec}}$ are initialized randomly, with no more optimization during the training. Only the output weights $\mathbf{W}^{\text{out}}$ are trained using linear regression. This results in the significantly simpler training procedure for RCN while the performance of the RCN has been shown to be comparable with more complex models [57, 58].

If U_t, R_t and Y_t represent the reservoir inputs, the reservoir outputs and the readouts at time t, the RCN equations can be written as follows:

$$R_t = (1-\lambda)R_{t-1} + \lambda f_{\text{res}}(\mathbf{W}^{\text{in}}U_t + \mathbf{W}^{\text{rec}}R_{t-1} + \mathbf{W}^{\text{b}}) \quad \text{(A1)}$$

$$Y_t = \mathbf{W}^{\text{out}}R_t \quad \text{(A2)}$$

with λ being a leaking rate between 0 and 1, with f_{res} being the nonlinear activation function of the reservoir neurons (we used *hyperbolic tangent* in this work) and with $\mathbf{W}^{\text{in}}$, $\mathbf{W}^{\text{rec}}$, $\mathbf{W}^{\text{b}}$ and $\mathbf{W}^{\text{out}}$ being the input, recurrent, bias and output weight matrices, respectively. Equation (A1) represents a leaky integration of the neuron activation and equation (A2) shows how Y_t is calculated based on the reservoir state R_t and the trained weight matrix $\mathbf{W}^{\text{out}}$.

For training, all reservoir states are collected in the reservoir state collection matrix $\mathbf{R}$. To add the intercept term for linear regression, every reservoir state R_t is expanded by a constant of 1. The desired outputs D_t, which are 0 for non-AEs and 1 for AE modes, are collected into the desired output collection vector $\mathbf{D}$. Afterwards, $\mathbf{W}^{\text{out}}$ is obtained using ridge regression, via equation (A3), to prevent overfitting to the training data. The regularization parameter $\epsilon = 0.01$ penalizes large values in $\mathbf{W}^{\text{out}}$, and $\mathbf{I}$ is the identity matrix. The size of the output weight matrix $N^{\text{out}} \times (N^{\text{res}} + 1)$ determines the total number of free parameters to be trained in RCNs.

$$\mathbf{W}^{\text{out}} = \left(\mathbf{R}\mathbf{R}^{\text{T}} + \epsilon\mathbf{I}\right)^{-1}(\mathbf{D}\mathbf{R}^{\text{T}}). \quad \text{(A3)}$$

In order to optimize the reservoirs main hyperparameters, we followed the instructions in [57, 64] which led to $(K^{\text{in}}, K^{\text{rec}}, \alpha_U, \alpha_R, \lambda) = (10, 10, 0.7, 0.9, 0.5)$.

ORCID iDs

Azarakhsh Jalalvand https://orcid.org/0000-0001-8739-1793
Alan A. Kaptanoglu https://orcid.org/0000-0002-6337-2907
Andrew O. Nelson https://orcid.org/0000-0002-9612-1936
Joseph Abbate https://orcid.org/0000-0002-5463-6552
Max E. Austin https://orcid.org/0000-0002-0017-8605
Geert Verdoolaege https://orcid.org/0000-0002-2640-4527
William W. Heidbrink https://orcid.org/0000-0002-6942-8043
Egemen Kolemen https://orcid.org/0000-0003-4212-3247

References

[1] Lang P. *et al* 2004 *Plasma Phys. Control. Fusion* **46** L31
[2] Ham C., Kirk A., Pamela S. and Wilson H. 2020 *Nat. Rev. Phys.* **2** 159–67
[3] Chen L. *et al* 2016 *Rev. Mod. Phys.* **88** 015008
[4] Todo Y. 2019 *Rev. Mod. Plasma Phys.* **3** 1–33
[5] Rea C. and Granetz R.S. 2018 *Fusion Sci. Technol.* **74** 89–100
[6] Fu Y. *et al* 2020 *Phys. Plasmas* **27** 022501
[7] Bosch H.-S. *et al* 2013 *Nucl. Fusion* **53** 126001
[8] Maljaars E. *et al* 2017 *Nucl. Fusion* **57** 126063
[9] Kolemen E. *et al* 2015 *J. Nucl. Mater.* **463** 1186–90
[10] Albanese R., Ambrosino R., Castaldo A., De Tommasi G., Luo Z.P., Mele A., Pironti A., Xiao B.J. and Yuan Q.P. 2017 *Nucl. Fusion* **57** 086039
[11] Morgan K.D., Hossack A.C., Hansen C.J., Nelson B.A. and Sutherland D.A. 2021 *Rev. Sci. Instrum.* **92** 053530
[12] Taylor R., Kutz J.N., Morgan K. and Nelson B.A. 2018 *Rev. Sci. Instrum.* **89** 053501
[13] Kaptanoglu A.A., Morgan K.D., Hansen C.J. and Brunton S.L. 2020 *Phys. Plasmas* **27** 032108
[14] Nayak I. *et al* 2020 Dynamic mode decomposition for prediction of kinetic plasma behavior *2020 Int. Applied Computational Electromagnetics Society Symposium (ACES)* (27-31 July, Virtual) pp 1–2
[15] Willcox K. and Peraire J. 2002 *AIAA J.* **40** 2323–30
[16] Ariola M., Ambrosino G., Pironti A., Lister J.B. and Vyas P. 2002 *IEEE Trans. Control Syst. Technol.* **10** 646–53
[17] Ariola M. *et al* 2005 *IEEE Control Syst. Mag.* **25** 65–75
[18] Moreau D. *et al* (Contributors to the EFDA-JET Workprogramme) 2003 *Nucl. Fusion* **43** 870
[19] Goodman T.P., Felici F., Sauter O. and Graves J.P. 2011 *Phys. Rev. Lett.* **106** 245002
[20] Levesque J.P. *et al* 2013 *Nucl. Fusion* **53** 073037
[21] Galperti C. *et al* 2014 *Plasma Phys. Control. Fusion* **56** 114012
[22] Galperti C., Coda S., Duval B.P., Llobet X., Milne P., Sauter O., Moret J.M. and Testa D. 2017 *IEEE Trans. Nucl. Sci.* **64** 1446–54
[23] Kaptanoglu A.A. *et al* 2021 *Phys. Rev.* E **104** 015206
[24] Kaptanoglu A.A. *et al* 2021 *Phys. Rev. Fluids* **6** 094401
[25] Goodfellow I. *et al* 2016 *Deep Learning* (Cambridge, MA: MIT Press) (https://deeplearningbook.org)
[26] Montes K.J. *et al* 2019 *Nucl. Fusion* **59** 096015
[27] Cannas B., de Vries P.C., Fanni A., Murari A., Pau A. and Sias G. 2015 *Plasma Phys. Control. Fusion* **57** 125003
[28] Rea C., Granetz R.S., Montes K., Tinguely R.A., Eidietis N., Hanson J.M. and Sammuli B. 2018 *Plasma Phys. Control. Fusion* **60** 084004
[29] Murari A., Lungaroni M., Peluso E., Gaudio P., Vega J., Dormido-Canto S., Baruzzo M. and Gelfusa M. 2018 *Nucl. Fusion* **58** 056002
[30] Kates-Harbeck J., Svyatkovskiy A. and Tang W. 2019 *Nature* **568** 526–31
[31] Bustos A. *et al* 2020 *Plasma Phys. Control. Fusion* **63** 095001
[32] Škvára V. *et al* 2020 *Fusion Sci. Technol.* **76** 962–71
[33] Woods B.J.Q., Duarte V.N., Fredrickson E.D., Gorelenkov N.N., Podesta M. and Vann R.G.L. 2020 *IEEE Trans. Plasma Sci.* **48** 71–81
[34] Guo B.H. *et al* 2020 *Plasma Phys. Control. Fusion* **63** 025008
[35] Turnbull A.D., Strait E.J., Heidbrink W.W., Chu M.S., Duong H.H., Greene J.M., Lao L.L., Taylor T.S. and Thompson S.J. 1993 *Phys. Fluids* B **5** 2546–53
[36] Heidbrink W.W., Strait E.J., Chu M.S. and Turnbull A.D. 1993 *Phys. Rev. Lett.* **71** 855
[37] Heidbrink W.W., Van Zeeland M.A., Austin M.E., Crocker N.A., Du X.D., McKee G.R. and Spong D.A. 2021 *Nucl. Fusion* **61** 066031
[38] Betti R. and Freidberg J.P. 1992 *Phys. Fluids* B **4** 1465–74
[39] Fasoli A. *et al* 1995 *Nucl. Fusion* **35** 1485
[40] Heidbrink W.W. *et al* 2020 *Nucl. Fusion* **61** 016029
[41] Sharapov S.E. *et al* 2002 *Phys. Plasmas* **9** 2027–36
[42] Kimura H. *et al* 1998 *Nucl. Fusion* **38** 1303
[43] Cheng C.Z., Chen L. and Chance M.S. 1985 *Ann. Phys., NY* **161** 21–47
[44] Cheng C.Z. and Chance M.S. 1986 *Phys. Fluids* **29** 3695–701
[45] Heidbrink W.W., Strait E.J., Doyle E., Sager G. and Snider R.T. 1991 *Nucl. Fusion* **31** 1635
[46] Wong K. *et al* 1991 *Phys. Rev. Lett.* **66** 1874–1877
[47] Heidbrink W.W. 2008 *Phys. Plasmas* **15** 055501
[48] Madsen B., Salewski M., Heidbrink W.W., Stagner L., Podestà M., Lin D., Garcia A. V., Hansen P.C. and Huang J. 2020 *Nucl. Fusion* **60** 066024
[49] Fu G.Y. 2008 *Phys. Rev. Lett.* **101** 185002
[50] Li P. *et al* 2021 *Nucl. Fusion* **61** 086020
[51] Tang S. *et al* 2021 *Phys. Rev. Lett.* **126** 155001
[52] Garcia-Munoz M. *et al* 2019 *Plasma Phys. Control. Fusion* **61** 054007
[53] Austin M.E. and Lohr J. 2003 *Rev. Sci. Instrum.* **74** 1457–9
[54] Bornatici M., Cano R., De Barbieri O. and Engelmann F. 1983 *Nucl. Fusion* **23** 1153
[55] Zaremba W. *et al* 2014 arXiv:1409.2329
[56] Jaeger H. 2001 The 'echo state' approach to analysing and training recurrent neural networks *Tech. Rep. GMD Report 148* German National Research Center for Information Technology (https://faculty.iu-bremen.de/hjaeger/pubs/EchoStatesTechRep.pdf)
[57] Jalalvand A., Abbate J., Conlin R., Verdoolaege G. and Kolemen E. 2021 *IEEE Trans. Neural Netw. Learn. Syst.* 1–12
[58] Jalalvand A. *et al* 2019 Radar signal processing for human identification by means of reservoir computing networks *IEEE Radar Confer. (RadarConf)* (22-26 April, Boston, USA) pp 1–6
[59] Steiner P. *et al* 2020 Multipitch tracking in music signals using echo state networks *28th European Signal Processing Conf. (EUSIPCO)* (18-21 January, Amsterdam, Netherlands) pp 126–30
[60] Ting K.M. 2010 *Encyclopedia of Machine Learning* ed C. Sammut *et al* (Berlin: Springer) p 209
[61] Acharya A.S. *et al* 2013 *India J. Med. specialities* **4** 330–3
[62] Nelson A.O., Logan N.C., Choi W., Strait E.J. and Kolemen E. 2020 *Plasma Phys. Control. Fusion* **62** 094002
[63] Triefenbach F. *et al* 2010 Phoneme recognition with large hierarchical reservoirs *Proc. NIPS* (6-12 December, Vancouver, Canada) pp 2307–15 (https://papers.nips.cc/paper/2010)
[64] Jalalvand A., Triefenbach F., Demuynck K. and Martens J.-P. 2015 *Comput. Speech Lang.* **30** 135–58
[65] Jalalvand A., Demuynck K., De Neve W. and Martens J.-P. 2018 *Neurocomputing* **277** 237–48
[66] Pathak J. *et al* 2018 *Phys. Rev. Lett.* **120** 024102