Explainable deep learning for the analysis of MHD spectrograms in nuclear fusion

IOP Publishing Mach. Learn.: Sci. Technol. 3 (2022) 015015 https://doi.org/10.1088/2632-2153/ac44aa

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
29 September 2021

REVISED
7 December 2021

ACCEPTED FOR PUBLICATION
20 December 2021

PUBLISHED
30 December 2021

PAPER

Explainable deep learning for the analysis of MHD spectrograms in nuclear fusion

Diogo R Ferreira[2,*], **Tiago A Martins**[2], **Paulo Rodrigues**[2] **and JET Contributors**[1,3]

[1] EUROfusion Consortium, JET, Culham Science Centre, Abingdon OX14 3DB, United Kingdom
[2] Instituto de Plasmas e Fusão Nuclear (IPFN), Instituto Superior Técnico (IST), Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
[3] See the author list of 'Overview of JET results for optimising ITER operation' by J Mailloux *et al* to be published in Nuclear Fusion special issue: overview summary papers from the 28th Fusion Energy Conference (Nice, France, 10–15 May 2021).
[*] Author to whom any correspondence should be addressed.

E-mail: diogo.ferreira@tecnico.ulisboa.pt

Keywords: interpretable machine learning, deep learning, convolutional neural networks, class activation mapping

Abstract

In the nuclear fusion community, there are many specialized techniques to analyze the data coming from a variety of diagnostics. One of such techniques is the use of spectrograms to analyze the magnetohydrodynamic (MHD) behavior of fusion plasmas. Physicists look at the spectrogram to identify the oscillation modes of the plasma, and to study instabilities that may lead to plasma disruptions. One of the major causes of disruptions occurs when an oscillation mode interacts with the wall, stops rotating, and becomes a locked mode. In this work, we use deep learning to predict the occurrence of locked modes from MHD spectrograms. In particular, we use a convolutional neural network with class activation mapping to pinpoint the exact behavior that the model thinks is responsible for the locked mode. Surprisingly, we find that, in general, the model explanation agrees quite well with the physical interpretation of the behavior observed in the spectrogram.

1. Introduction

Nuclear fusion is the energy-generating process that powers the stars, and that several public-funded initiatives—as well as a growing number of startups—are trying to replicate using a variety of approaches, from small-scale to large-scale devices, namely ITER [1]. One of the most promising configurations is the tokamak [2], a toroidal chamber where the plasma (which is basically a compound of hydrogen isotopes heated up to about 150 million degrees Celsius) is confined by a combination of magnetic fields.

Despite having been proposed more than half a century ago, the concept of tokamak is still a rich source of technical and scientific challenges, because the plasma, when confined in a toroidal geometry, behaves in intricate ways and is subject to a number of possible instabilities [3]. These instabilities can be studied by considering the plasma as a fluid (an overall charge-neutral, but electrically conductive fluid) immersed in a magnetic field, which is described by the theory of magnetohydrodynamics (MHDs) [4].

The MHD behavior of a plasma can be probed through the use of magnetic pick-up coils, also known as Mirnov coils [5], which detect amplitude fluctuations in the magnetic field as the plasma rotates inside the torus. Of special interest is the frequency of such fluctuations, which can be used to determine the oscillation modes of the plasma [6]. Some of these modes are known to be detrimental to the plasma confinement, and can even cause major disruptions [7]. These are catastrophic events, where the plasma confinement is suddenly lost, with severe consequences for plasma-facing materials.

Therefore, analyzing the MHD behavior of a plasma, as provided by the signals of magnetic pick-up coils, is of utmost importance to understand the physics of a fusion experiment and, in particular, to identify any behavior that could potentially throw the experiment off course, and into a disruption. Such analysis is typically carried out by means of MHD spectrograms which, in a similar way to audio spectrograms, plot the frequency spectrum of a signal at each point in time, based on a windowed short-time Fourier transform [8].

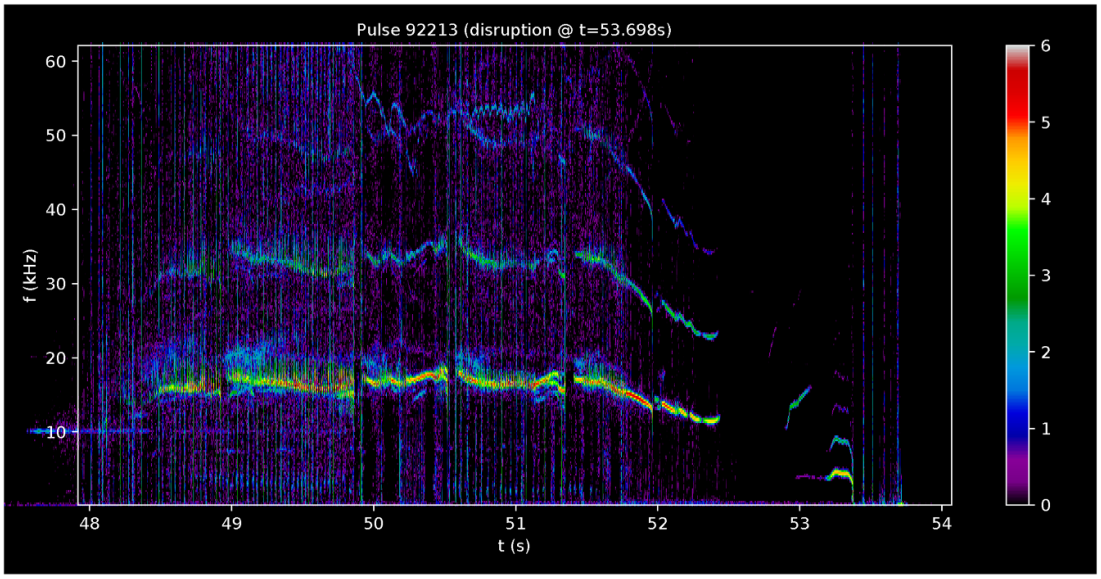

Figure 1. MHD spectrogram for JET pulse 92 213 from $t = 48$ s to $t = 54$ s. The spectrogram is plotted with logarithmic amplitude, so the color intensities have arbitrary units.

As a concrete example, figure 1 shows the MHD spectrogram for a sample pulse at the Joint European Torus (JET), which is currently the largest operating tokamak, before ITER. The spectrogram highlights the frequencies of the magnetic fluctuations through time, up to the end of the pulse. In this particular case, the pulse ends with a disruption at around $t = 53.7$ s.

A trained physicist will use this and other tools to find the oscillation modes that were active during the pulse, from about $t = 48.5$ s to $t = 52.5$ s. It is visible that the plasma starts decelerating at around $t = 51.5$ s (as the heating systems are being turned off, the plasma loses momentum) and that there is an interruption of the MHD activity at around $t = 52.5$ s. When the activity resumes, after $t = 53.0$ s, the plasma is rotating much slower, and the modes are brought to a sudden halt at around $t = 53.4$ s, which eventually causes a major disruption.

This last phenomenon is called mode locking [9] and it is due to the fact that the plasma oscillations induce eddy currents in the surrounding structures, which (in the absence of further heating) contribute to decelerate it even further, until the mode stops rotating and becomes locked, causing a loss of confinement. In fact, the locked mode is one of the strongest indicators (but not the only one) that a disruption is about to happen, and it has been used as the single most predictive feature in several disruption prediction studies, e.g. [10, 11].

However, more recently it has been realized that it is necessary to go beyond the locked mode, because: (1) it happens too close to the disruption to be able to change the course of events, and (2) it does not explain what happened before that, i.e. what actually caused the locked mode to occur and the disruption to ensue. It is in this second part that we bring deep learning into the picture, to try to predict the occurrence of a locked mode based on the MHD activity that has been observed up to that point in the pulse.

Moreover, we are not content just by predicting locked-mode events; we want to know exactly which MHD activity makes the model believe that there will be a locked mode. Therefore, we use a localization technique—specifically, class activation mapping (CAM) [12]—to identify the regions in the spectrogram that contribute the most, either positively or negatively, to the model prediction.

Before we start describing our model, it is important to note that this is not the first time that deep learning has been applied to MHD spectrograms, or that CAM has been applied to fusion data. In a recent paper, Bustos *et al* [13] used deep learning to identify, via segmentation, the active oscillation modes in MHD spectrograms; and Kwon *et al* [14] used CAM over images from a video camera to classify the observed behavior as disruptive or non-disruptive, and to locate the area in the image that affected the classification.

In any case, to the best of our knowledge, this is the first time that CAM is being used on MHD spectrograms for the purpose of predicting the occurrence of the locked mode and to identify the MHD behavior that is responsible for that occurrence. In the next section, we show that there is an intimate relationship between the locked mode and disruptions. In subsequent sections, we describe how we trained a convolutional neural network (CNN) to predict the locked mode, and how we used CAM to localize the relevant behavior.

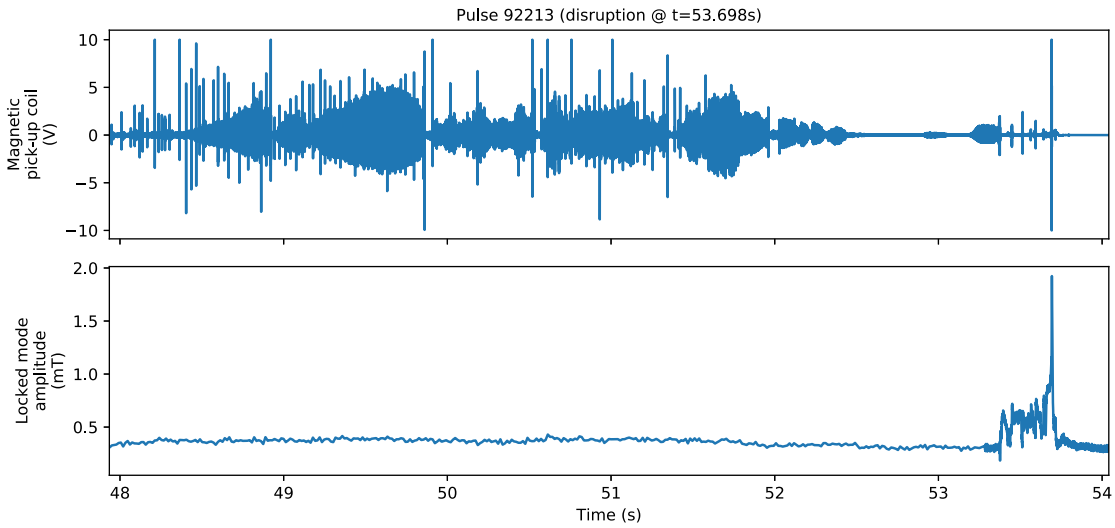

Figure 2. Magnetic pick-up coil (*top*) and locked mode amplitude (*bottom*) signals for JET pulse 92 213 from $t = 48$ s to $t = 54$ s.

2. Locked mode and disruptions

Besides the magnetic pick-up coils that provide the signal to generate the spectrogram, the magnetic diagnostics at JET include many other sensors, namely to measure the locked mode amplitude. Figure 2 shows, at the top, the signal from the magnetic pick-up coil that is used to generate the spectrogram and, at the bottom, the measured locked mode amplitude.

It is apparent that there is an increase in the locked mode amplitude at around $t = 53.4$ s, the same time where we have seen the plasma rotation coming to a stop in the MHD spectrogram (figure 1). We also know that this pulse ended with a disruption, which can be seen as a spike in both signals at $t = 53.7$ s.

The question now is whether it is possible to establish a direct relationship between the increase in the locked mode amplitude and the fact that the pulse ends with a disruption. For this purpose, and similarly to what others have done [15], we compared the locked mode amplitude of disruptive and non-disruptive pulses, and tried to find a threshold to separate them (i.e. a threshold on the locked mode amplitude that, in general, is exceed by disruptive pulses but is not exceed by non-disruptive pulses). Such threshold can be regarded as a simple binary classifier that works as follows:

- If the pulse is disruptive and the locked mode amplitude exceeds the threshold before the disruption time, then the pulse is classified as a true positive.
- If the pulse is non-disruptive and the locked mode amplitude does not exceed the threshold at any point in time, then the pulse is classified as a true negative.
- If the pulse is non-disruptive and the locked mode amplitude exceeds the threshold at any point in time, then the pulse is classified as a false positive. In the fusion community, this is also referred to as a false alarm.
- If the pulse is disruptive and the locked mode amplitude does not exceed the threshold at any point in time or, if it exceeds, it does so on or after the disruption time, then the pulse is classified as a false negative. In the fusion community, this is also referred to as a missed alarm.

Training this classifier is equivalent to finding the optimal threshold that maximizes some given metric, such as accuracy, balanced accuracy, F1 score, or Matthews correlation coefficient [16], for example. Figure 3 shows that all of these metrics are maximized when the threshold value is 1.069 mT. (For comparison, the threshold that is used for machine protection at JET is 2 mT [17].) At that point, the accuracy is 97.9%, the balanced accuracy is 97.6%, the F1 score is 97.4%, and the Matthews correlation coefficient is 95.7%; the false alarm rate is 1% and the missed alarm rate is 2.5%.

The fact that all of these metrics are above 95% shows that the locked mode is an excellent discriminator between disruptive and non-disruptive pulses. This study was conducted over a dataset of 485 pulses from the JET baseline scenario [18], which is the operational scenario that is being developed for ITER. In this dataset, there were 39% disruptive pulses and 61% non-disruptive ones, so the dataset was fairly balanced, which explains the agreement between accuracy and balanced accuracy, and across all metrics in general. We will be using this same dataset in the forthcoming sections.

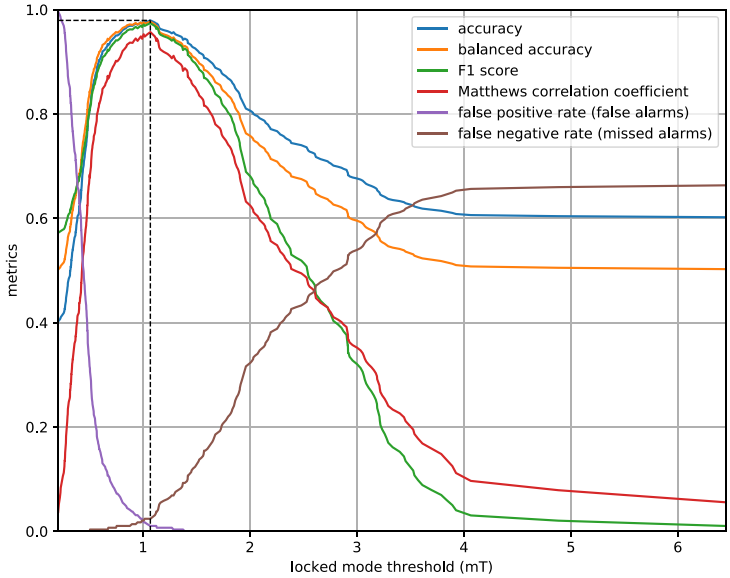

Figure 3. Binary classification metrics for disruption prediction with a threshold on the locked mode amplitude, for a dataset of 485 baseline pulses.

3. Deep learning model

In the previous section, we have shown that the locked mode is highly correlated with the occurrence of disruptions. However, in a general setting, not all disruptions are due to MHD activity or locked modes [19], so it does not make perfect sense to predict disruptions directly from MHD spectrograms. What makes sense is to predict the development of a locked mode from MHD activity, since locked modes are essentially MHD-related phenomena.

Therefore, we use the results from the previous section to build a binary classifier, in the form of a deep learning model, that receives a segment of MHD activity and classifies that segment as coming from a pulse that exceeds (class 1) or does not exceed (class 0) the threshold of 1.069 mT on the locked mode amplitude. The goal is to predict whether a pulse with the given MHD activity will develop a locked mode or not.

The model that we devised for this task is a CNN that performs binary classification over an input image. The input image is a window taken from an MHD spectrogram, where the horizontal axis represents time and the vertical axis represents frequency. For practical reasons, we try to have the same resolution on both axes, as will be explained below. Meanwhile, the model architecture can be readily seen in figure 4.

3.1. Model input

The model receives a 256×256 image that is obtained as follows:

- The MHD spectrogram is computed from a downsampled version of the raw signal coming from the magnetic pick-up coil. Although the signal comes from a fast acquisition system that provides a sampling rate of up to 2 MHz, for the purpose of this work we downsampled it to 125 kHz, which is enough for the range of frequencies that need to be analyzed, and is comparable to the downsampling that is usually performed when applying deep learning to high temporal resolution data from fusion diagnostics [20].
- We then apply a short-time Fourier transform with a length of 512 points and a step of 512 points as well. For a 125 kHz signal, a step of 512 points means that the spectrogram has a time resolution of 4.096 ms, and a length of 512 points means that there are 256 points on the positive side of the frequency axis. Since the time scale of the phenomena that we want to analyze is on the order of a second [21], we use a window of 256×256 over the spectrogram. On the time axis, this window covers 1.04858 s, and on the frequency axis it covers the whole vertical range of the spectrogram.

The input image that is provided to the model can be sampled anywhere on the time axis of the spectrogram. In particular, it may be sampled at the very beginning of the pulse, where there is no indication about the future outcome of the pulse; or it may be sampled past the disruption in a disruptive pulse, where the locked mode has already happened. For the sake of generality, and in order to avoid biasing the model toward a desired outcome, we allow the 256×256 window to be sampled anywhere.

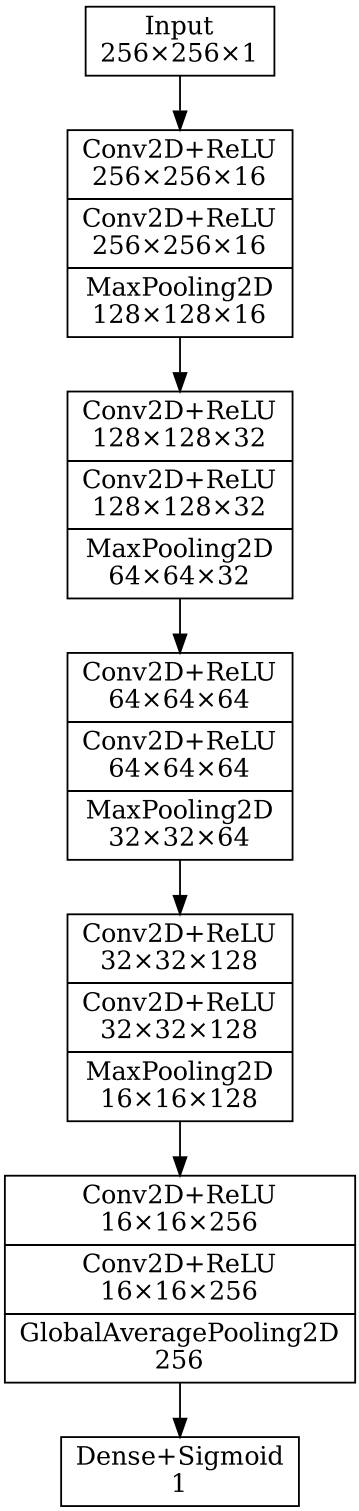

Figure 4. CNN to predict the occurrence of a locked mode from a 256 × 256 image drawn from an MHD spectrogam.

3.2. Model structure

The model is composed for five convolutional blocks. Each block has two convolutional layers followed by a max-pooling layer, except for the last block which introduces a global average pooling layer at the end. This has been done on purpose, to facilitate the generation of the class activation maps later on. However, there is no reason to worry about this having a detrimental effect in the results, since our experiments suggest that a classical CNN with max-pooling and two dense layers at the end achieves essentially the same results, because both models have more than enough parameters to overfit the training data. In fact, the use of global average pooling has the added benefit of making the model less prone to overfitting [22].

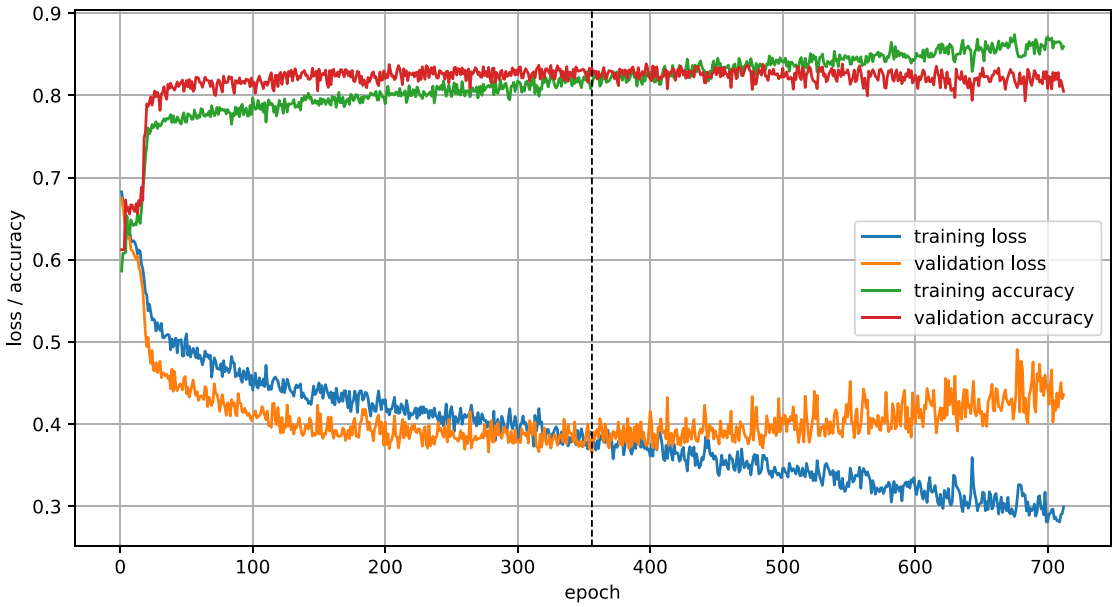

Figure 5. Loss and accuracy during the training process. The dash line indicates the point at which the minimum validation loss was achieved.

3.3. Model training

The model was trained by drawing samples from the MHD spectrograms of 485 baseline pulses, where 90% (436 pulses) were used for training, and 10% (49 pulses) were used for validation. The data splitting was done in a such a way as to guarantee that the distribution of disruptive and non-disruptive pulses (39% vs. 61%, respectively) was the same in both the training set and the validation set.

The batching process was such that each training batch contained exactly one sample from each training pulse; so the batch size was equal to the number of training pulses. However, for each batch a different sample is drawn from the MHD spectrogram of each training pulse. As for the validation pulses, we took 100 evenly-spaced samples from each of them, and used those samples for validation at the end of each training epoch. Each epoch consisted in providing the model with ten training batches.

Although it is possible to draw many 1 s samples from the spectrogram of any given pulse, the fact that the number of pulses is limited (and hence the variety of the observed behaviors) has led us to use all the available pulses for training/validation purposes, with the exception a single pulse (92 213) that was left out for testing purposes. This is a pulse that is particularly familiar and representative of a class of disruptions whose dynamics have been studied before [23].

Figure 5 shows the evolution of the loss (binary cross-entropy) and classification accuracy (percentage) during training. In general, and across many trials with different hyperparameters, the validation loss hardly falls below 0.4, and the validation accuracy does not rise much above 80%, which, nevertheless, is considered to be a good result, especially when taking into account that the training samples can be drawn from anywhere in the MHD spectrograms.

In the training run of figure 5, the best validation loss (0.365) is achieved on epoch 356; after that, it tends to increase in a clear sign of overfitting. At that point (epoch 356), the validation accuracy achieves 83.3%. The whole training process is rather noisy due to the fact that the model keeps receiving ambiguous samples. Specifically, at the beginning of a pulse it is impossible to say whether the pulse will develop a locked mode or not, because the behavior is similar in both disruptive and non-disruptive pulses, but the model receives a sample labeled 0 or 1, and has to adjust accordingly. This often leads to spikes that are especially visible in the validation loss.

A different labeling scheme could help avoid this jitter in the training process, and possibly provide even higher accuracy. For example, a sample could be labeled as 1 only if it is within a certain time interval from the locked mode. However, this would introduce additional problems, such as having to decide on an appropriate time interval, and it would also contribute to make the dataset severely unbalanced by reducing the number of samples that are labeled as 1, in comparison to those that are labeled as 0.

3.4. Model predictions

After training the model, we can use it to make predictions on pulses that the model has not seen before. For this purpose, we can slide a window across the time axis of the spectrogram, from the beginning to the end of

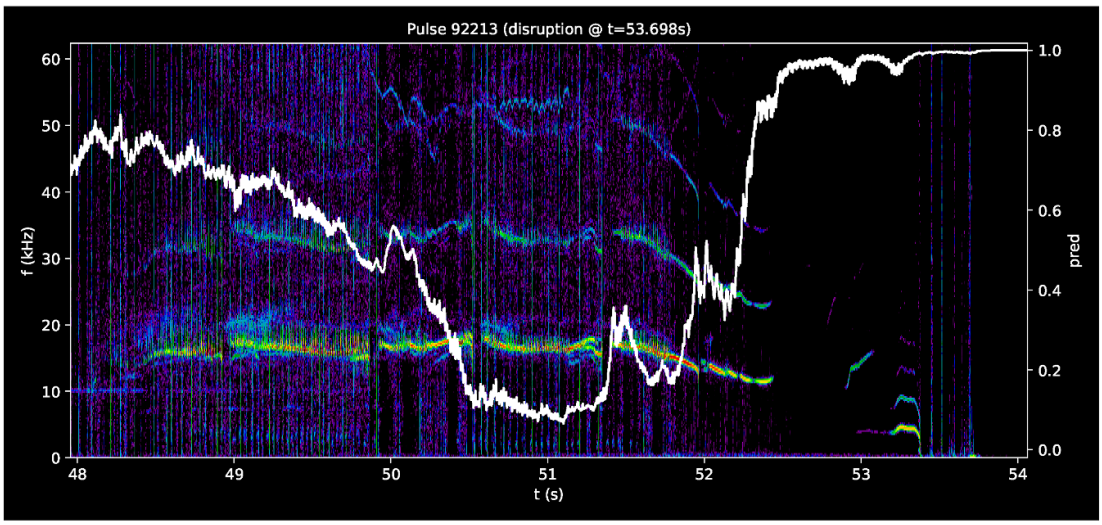

Figure 6. Model prediction at each point in time for pulse 92 213, overlaid on the original spectrogram for that pulse.

the pulse, and ask the model for a prediction at each point in time, where the window is positioned always to the left (i.e. to the past) of the point in time for which the prediction is being made. Figure 6 shows the result for pulse 92 213.

It is interesting to note that the prediction starts relatively high at around $t = 48$ s because the model is confused by a numerical artifact that exists at the beginning of this spectrogram, at 10 kHz. The prediction eventually drops down to low values around $t = 51$ s before rising up again when the plasma starts decelerating, and it shoots up when the MHD activity is interrupted at $t = 52.5$ s. From that point onwards, the prediction is very close to 1.0 and, once the model sees the disruption, it becomes absolutely certain that this pulse has a locked mode.

The most interesting part of this prediction is that, for a pulse that the model has never seen before, the model is pretty certain that this pulse will develop a locked mode well before it happens. For this type of pulses, the warning time is typically in the range of 1.0–1.5 s before the disruption, which is well above the average warning time of other disruption predictors, such as APODIS [24], which relies mainly on the locked mode amplitude signal and has an average warning time of 350 ms. The inference time of the model is about 4 ms on a single GPU, including the latency associated with data transfers to and from GPU memory.

In the next section, we use CAM to understand what makes the model become so certain about that prediction.

4. Model interpretability

CAM [12] is a localization technique that provides a heatmap highlighting the regions of an input image that most contributed to the classification of that image by the model. Originally, CAM has been devised for CNNs with a global average pooling layer, as in our model in figure 4.

Using our own model to illustrate the concept, the last convolutional layer produces 256 features maps of size 16×16. What the global average pooling layer does is to reduce each of those features maps to an average value. The 256 average values that result from such reduction are then passed to a dense layer, which has its own weights to produce the model prediction.

If we go back to the feature maps we have before the global average pooling, and assign the weights of the dense layer to those corresponding feature maps, and then do a weighted average of the feature maps, we end up with a single 16×16 result. This result can be interpreted as a heatmap where the value of each pixel indicates how important that pixel is for the classification of the image that was given as input to the model.

There is, however, a slight caveat: while the input image is 256×256, the heatmap is only 16×16. The solution, which is part of the original CAM approach, is to scale up the heatmap to the same dimensions of the input image. The heatmap can then be overlaid on top of the input image to highlight the regions that most contributed to the classification.

4.1. Changing the model

The CAM approach requires a few adaptations to the model. Once the model has been trained, we remove the global average pooling layer, and also retrieve the weights from the dense layer. We replace those two

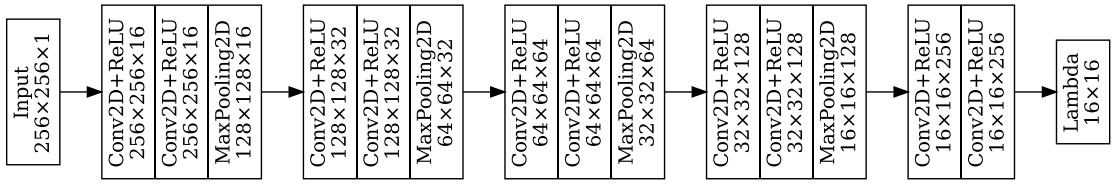

Figure 7. Model used to generate the class activation map for a given input image.

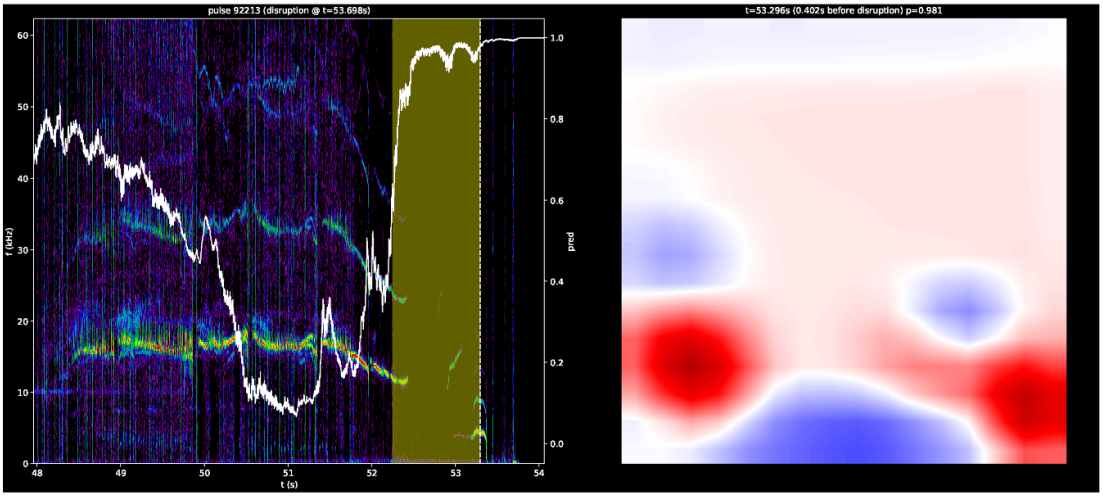

Figure 8. Spectrogram, prediction and class activation map for pulse 92 213 at $t = 53.3$ s.

layers with a single custom layer that computes the weighted average of the feature maps produced by the last convolutional layer. The resulting model is shown in figure 7.

4.2. Analyzing the class activation maps

The class activation maps produced by the model in figure 7 turn out to be a very useful tool to understand how the model in figure 4 arrives at a certain prediction. In many cases, the regions highlighted in the class activation map agree with the physical intuition of what is going on with the pulse at that point in time. Figure 8 shows an example.

On the left-hand side of figure 8, there is the spectrogram and the model predictions, as shown earlier in figure 6. In addition, there is a yellow rectangular region over the spectrogram, highlighting the 256×256 window that the model is currently looking at. (The window appears to be rectangular rather than square because the time axis is somewhat compressed with respect to the frequency axis.) It is based on this window that the model produces the prediction at the time point indicated by a vertical dash line on the rightmost edge of that window (i.e. at the first time point immediately after that window).

On the right-hand side of figure 8, there is the class activation map that corresponds to the yellow region on the left-hand side of the figure. For this class activation map, we chose a color scheme according to the following ideas:

- White is used to indicate neutral regions that neither contribute to classify the pulse into class 1 nor contribute to classify the pulse into class 0. The value on these pixels is zero (or very close to zero), which would yield a value of ~ 0.5 should they be passed through a sigmoid (without bias).
- Red is used to indicate positive regions that contribute to classify the pulse into class 1. The more intense red indicates higher values that might easily saturate a sigmoid toward 1.0.
- Blue is used to indicate negative regions that contribute to classify the pulse into class 0. The more intense blue indicates lower values that might easily saturate a sigmoid toward 0.0.

An interesting conclusion that one can draw from here is that the model prediction is a combination of all these contributions, and while the blue regions contribute toward class 0, the red regions contribute towards class 1. As a consequence, a prediction of 0.5 does not mean that the class activation map is white all

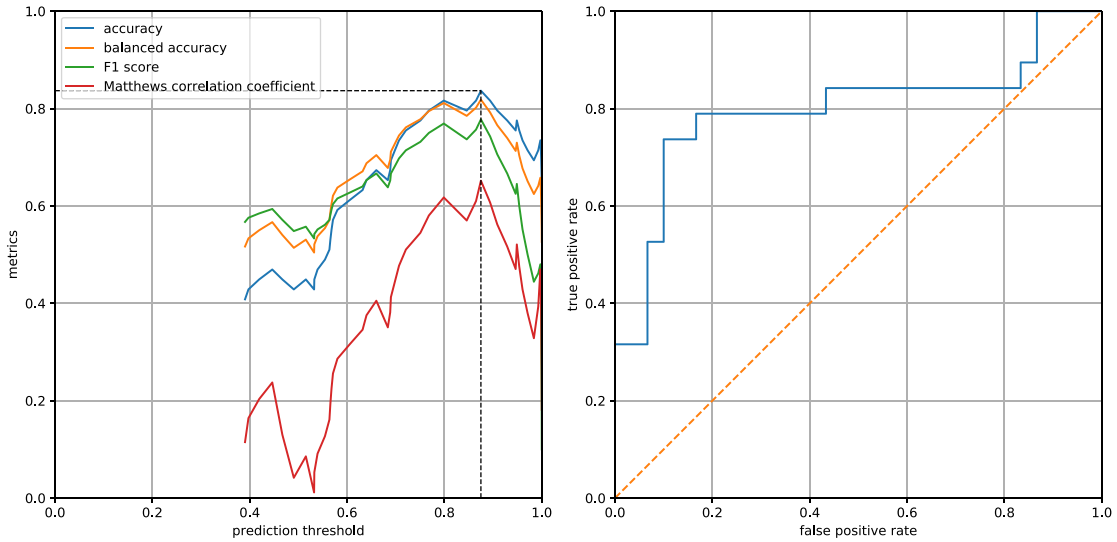

Figure 9. Classification metrics and ROC curve for a variable threshold on model output. Results are shown for the 10% of pulses in the validation set (49 pulses).

over; it just means that there is balance between the blue regions and the red ones. In the example of figure 8, the prediction is 0.981 but there are still some blue regions in the class activation map.

What is most striking in figure 8 is that the red regions in the class activation map highlight features that are very meaningful from a physics point of view: first the interruption of MHD activity at $t = 52.5$ s, and then its resumption at $t = 53.2$ s, before the locked mode develops. In particular, the model is paying special attention to that strongest, lowest frequency mode, which has been identified in the literature as a 2/1 mode that is highly susceptible of becoming locked [25].

Between the two red regions, there is a blue region which highlights the fact that, in the absence of MHD activity, there is no reason to believe that the pulse will develop a locked mode. However, this does not prevent the model from finding strong indications that a locked mode is about to occur, as suggested by the model prediction, which is consistently above 0.9 in this period. (Naturally, if the window was shorter in time, this would prevent the model from seeing the whole picture; this highlights the need for sufficiently long windows and the importance of time-evolution patterns.)

A legitimate question that one may ask is how large does the model prediction need to become for us to seriously consider the possibility that a locked mode will develop? In other words, what is the threshold above which we should throw an alarm? We address this question in the next subsection.

4.3. Using a threshold on the model output

When training the model, the pulses were split into 90% for training and 10% for validation, with only one pulse being left out for testing purposes. This means that we do not have a proper test set on which to evaluate the model. Of course, there are many other pulses at JET on which we could test our model, but these belong to different experimental scenarios where the model might not perform so well. Evaluating the model on the training data is out of the question, but perhaps we can turn to the validation set to extract some metrics.

Figure 9 shows what happens when we look at the model predictions on the validation set and we use a threshold to decide when to throw an alarm. As in section 2, throwing an alarm after the locked mode has happened is the same a not throwing an alarm (false negative or missed alarm), and throwing an alarm on a pulse without locked mode is a false positive or false alarm.

Figure 9 shows that all the metrics (namely accuracy, balanced accuracy, F1 score, and Matthews correlation coefficient) are maximized at a threshold value of about 0.88. This means that, for the example shown earlier in figure 8, an alarm would have been thrown immediately upon the interruption of MHD activity at $t = 52.5$ s. Figure 9 also plots the receiver operating characteristic curve [26], and the area under this curve (AUC) is about 0.80, which is fairly good for this predictor. The fact that the AUC is well above 0.5 means that the model is able to recognize true positives (i.e. actual occurrences of the locked mode) without incurring into an equal proportion of false positives (i.e. false alarms).

5. Conclusion

In this work, we have reached three main findings.

First, there is an intimate relationship between the occurrence of locked modes and disruptions. This has been already reported at length in the literature, but it is comforting to see that it applies to our dataset as well. The fact that the locked mode is usually present when there is a disruption is what justifies and motivates the idea of predicting disruptions from MHD activity. In this work we do not predict disruptions directly but predict the occurrence of locked modes which, almost invariably, lead to disruptions.

Second, a deep learning model applied over MHD spectrograms is able to capture the behavioral patterns that lead to the locked mode, and is able to anticipate the locked mode with an advance warning time that is comparable to, or even exceeds, current disruption predictors. It should be noted that most of these predictors use the locked mode amplitude as one of their input signals, so analyzing the MHD spectrogram seems to provide more information than the locked mode signal alone, as it should be expected.

Third, CAM is an effective technique to understand and explain the predictions of a deep learning model. In our context, the use of CAM made it possible to extract some key insights that physicists are well aware of when analyzing MHD spectrograms. This means that when we try to explain why the model came up with a certain prediction, the explanation makes sense from a physics point of view. This explanation can be used as a starting point for a deeper investigation MHD phenomena.

In future work, we plan to extend this methodology to other experimental scenarios at JET, where the range of MHD phenomena may require the use of a larger model, with more learning capacity. The parameter and hyperparameter spaces for the current model could be more extensively explored, in order to possibly improve the results. The model could be incorporated as an additional alarm trigger in the JET real-time network. As an MHD analysis tool, the model could incorporate other interpretability techniques, which could bring additional insights.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgments

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014–2018 and 2019–2020 under Grant Agreement No. 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission. IPFN (Instituto de Plasmas e Fusão Nuclear) received financial support from FCT (Fundação para a Ciência e a Tecnologia) through projects UIDB/50010/2020 and UIDP/50010/2020.

ORCID iDs

Diogo R Ferreira https://orcid.org/0000-0001-5818-9406
Tiago A Martins https://orcid.org/0000-0002-4328-1151
Paulo Rodrigues https://orcid.org/0000-0001-6189-6865

References

[1] Claessens M 2020 *ITER: The Giant Fusion Reactor* (Berlin: Springer)
[2] Wesson J 2011 *Tokamaks* (Oxford : Oxford University Press)
[3] Zohm H 2015 *Magnetohydrodynamic Stability of Tokamaks* (Weinheim: Wiley-VCH)
[4] Freidberg J P 2014 *Ideal MHD* (Cambridge: Cambridge University Press)
[5] Igochine V 2015 *Active Control of Magneto-Hydrodynamic Instabilities in Hot Plasmas* (Berlin: Springer)
[6] Kim J S, Edgell D H, Greene J M, Strait E J and Chance M S 1999 *Plasma Phys. Control. Fusion* **41** 1399–420
[7] Bondeson A, Parker R, Hugon M and Smeulders P 1991 *Nucl. Fusion* **31** 1695–716
[8] Bizarro J P S and Figueiredo A C A 2008 *Fusion Eng. Des.* **83** 350–3
[9] Nave M and Wesson J 1990 *Nucl. Fusion* **30** 2575–83
[10] Vega J, Moreno R, Pereira A, Dormido-Canto S and Murari A 2015 Advanced disruption predictor based on the locked mode signal: application to JET *1st Conf. on Plasma Diagnostics*
[11] Sias G, Cannas B, Fanni A, Murari A and Pau A 2019 *Fusion Eng. Des.* **138** 254–66
[12] Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A 2016 Learning deep features for discriminative localization *2016 Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 2921–9
[13] Bustos A, Ascasíbar E, Cappa A and Mayo-García R 2021 *Plasma Phys. Control. Fusion* **63** 095001
[14] Kwon G, Wi H and Hong J 2021 *Fusion Eng. Des.* **168** 112375

[15] Gerasimov S et al 2019 Locked mode and disruptions in JET-ILW 46th Conf. on Plasma Physics
[16] Chicco D and Jurman G 2020 BMC Genomics 21 6
[17] Reux C et al 2013 Fusion Eng. Des. 88 1101–4
[18] Garzotti L et al 2019 Nucl. Fusion 59 076037
[19] de Vries P, Johnson M, Alper B, Buratti P, Hender T, Koslowski H and Riccardo V 2011 Nucl. Fusion 51 053018
[20] Churchill R M, Tobias B and Zhu Y 2020 Phys. Plasmas 27 062510
[21] Boozer A H 2012 Phys. Plasmas 19 058101
[22] Lin M, Chen Q and Yan S 2014 Network in network 2nd Int. Conf. on Learning Representations (ICLR)
[23] Ferreira D R, Carvalho P J, Carvalho I S, Stuart C, Lomas P J and JET Contributors 2021 Fusion Eng. Des. 164 112179
[24] Moreno R, Vega J, Dormido-Canto S, Pereira A, Murari A and JET Contributors 2016 Fusion Sci. Technol. 69 485–94
[25] Sweeney R, Choi W, Haye R L, Mao S, Olofsson K and Volpe F V 2016 Nucl. Fusion 57 016019
[26] Bradley A P 1997 Pattern Recognit. 30 1145–59