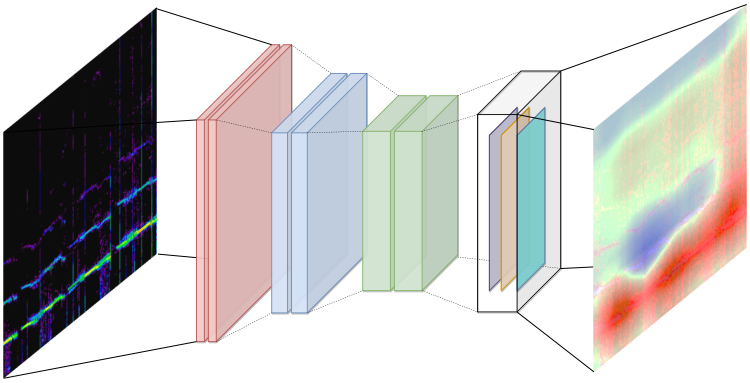
TÉCNICO LISBOA



Use of MHD Activity for Disruption Prediction in Tokamaks

Tiago Agostinho Martins

Thesis to obtain the Master of Science Degree in

Engineering Physics

Supervisors: Prof. Diogo Manuel Ribeiro Ferreira
Dr. Paulo Jorge Rodrigues

Examination Committee

Chairperson: Prof. Luís Paulo Da Mota Capitão Lemos Alves
Supervisor: Prof. Diogo Manuel Ribeiro Ferreira
Member of the Committee: Prof. Rui Miguel Dias Alves Coelho

October 2021

Declaração
Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Declaration
I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

- “If I have seen further, it is by standing on the shoulders of Giants”

Isaac Newton

Acknowledgments

To begin with, I must thank my supervisors, Professor Diogo Ferreira and Professor Paulo Rodrigues, for their continuous insight and guidance throughout the development of this thesis, which included crucial advises in many stages, from code development to theoretical comprehension. Their knowledge was also important to improve my capabilities in both nuclear fusion and deep learning domains.

I also want to send a word of appreciation to the JET team for the data access and their feedback towards the topic of this work. Of equal importance was the support provided by the IPFN team in the delivery of the necessary hardware, for which I am thankful.

This work marks the culmination of a 6 year journey that would not have been possible to complete without the support I received in this period from many people, and to which I am forever grateful. To my father and mother, Carlos and Miquelina, for always being present when I needed the most and for making me the person I am today, giving me all their support to pursue my goals and dreams. To my sister, Vanessa, for the love and for being my academic role model. To the rest of my family, for their unconditional support and kindness. And finally, to all my colleagues and friends for their friendship and encouragement, even in the most difficult times. A note of acknowledgment to Miguel Bento, Joana Simões, Ricardo Costa, João Rodrigues, Vera Algarvio, Rodrigo Antunes, António Coelho, Beatriz Pereira, Abhishek Gupta, and all of my colleagues from FCT and IST. Also to Fábio Moniz and Óscar Amaro for their availability to proofread this document. To all of them, my deepest gratitude.

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. 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. The author is thankful for the granted use of computational resources at CCFE (Culham Centre for Fusion Energy) in the UK, as well as to NVIDIA Corporation for the donation of hardware that was used to this project.

Resumo

As disrupções em tokamaks são atualmente um dos maiores desafios para a viabilidade da fusão nuclear e a sua validação experimental. Uma das principais causas de disrupções deve-se à interacção de certos modos instáveis de magnetohidrodinâmica (MHD) com a parede do reactor, causando a sua interrupção e paragem no referencial da máquina - modos locked (modos bloqueados). Estes modos são extensamente reconhecidos pela comunidade como desencadeadores de disrupções. Neste trabalho, pretendemos utilizar ferramentas de aprendizagem profunda para prever este tipo de disrupções e perceber possíveis mecanismos desencadeadores.

Numa primeira fase, é treinado um modelo baseado numa rede neuronal convolucional (CNN) para receber atividade MHD de uma experiência em forma de espectrograma e prever se a mesma é disruptiva devido a modos locked. O modelo consegue distinguir razoavelmente as classes de classificação, embora não possa ser diretamente comparado com outros classificadores de estado-da-arte.

Posteriormente, uma ferramenta de interpretabilidade em aprendizagem automática para redes neuronais convolucionais, denominada mapas de ativação de classe (CAM), é utilizada na tentativa de perceber qual a atividade MHD relevante que o modelo considera para a classificação e verificar a sua validação com a intuição física.

Os resultados indicam um foco considerável do método CAM para a interrupção de atividade MHD, em particular modos kink internos, e a sua resurgência antes do desenvolvimento dos modos locked, sendo congruente com a explicação física dada pelo espectrograma e com estudos anteriores.

Palavras-chave

Tokamaks; Magnetohidrodinâmica; Previsão de disrupções; Espectrogramas; Redes neuronais convolucionais; Aprendizagem automática interpretável

Abstract

Disruptions in tokamaks are a crucial challenge for the viability of nuclear fusion and its experimental validation. One of the main causes of disruptions is due to the interaction of certain unstable magnetohydrodynamic (MHD) modes of the plasma with the reactor wall, where they stop in the machine's reference frame - locked modes. These modes are widely recognized by the community as disruption triggers. In this work, we intend to use deep learning tools to predict these types of disruptions and to understand other possible triggering mechanisms.

First, a model based on a convolutional neural network (CNN) is trained to receive MHD activity from an experiment in the form of a spectrogram and to predict whether it belongs to a discharge ending in disruption due to mode-locking. The model can reasonably distinguish the classification classes, although it cannot be compared directly with other state-of-the-art predictors.

Subsequently, an interpretable machine learning tool for convolutional neural networks, called class activation mapping (CAM), is used in an attempt to understand which MHD activity is relevant to the model for classification and to verify its validity with the physical intuition.

The results indicate a considerable focus placed by the CAM method in the interruption of MHD activity, in particular internal kink modes, followed by its resurgence before the development of the locked mode. This observation is congruent with the physical explanation given by the spectrogram and with previous studies.

Keywords

Tokamaks; Magnetohydrodynamics; Disruption prediction; Spectrograms; Convolutional neural networks; Interpretable machine learning

Contents

List of Tables

List of Figures

Acronyms

APODIS	Advanced Predictor Of Disruptions
AUC	Area under the ROC curve
CAM	Class Activation Mapping
CNN	Convolutional Neural Networks
DMS	Disruption Mitigation System
DMV	Disruption Mitigation Valve
ECE	Electron Cyclotron Emission
FFT	Fast Fourier Transform
FRNN	Fusion Recurrent Neural Networks
ICRH	Ion Cyclotron Resonance Heating
ILW	ITER-like Wall
ITER	International Thermonuclear Experimental Reactor
JET	Joint European Torus
MGI	Massive gas injection
MHD	Magnetohydrodynamics
NBI	Neutral-beam Injection
NTM	Neoclassical Tearing Modes
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Networks
ROC	Receiver operating characteristic curve
RWM	Resistive Wall Modes
SPAD	Single signal Predictor based on Anomaly Detection
SPI	Shattered injected pellets
STFT	Short-time Fourier Transform
SVM	Support-vector machines
TTD	Time-to-disruption

1

Introduction

Contents

Nuclear fusion has been one of the main candidates for replacing conventional methods of electricity production. The environmental impact caused by the use of fossil fuels and the lack of congruence in society regarding the use of nuclear fission-based power plants reinforces the position of fusion as the Holy Grail of energy. The study of the Sun and its thermonuclear reaction mechanisms marked a starting point for initial fusion research. Among the various contributions, the works from Arthur Eddington [1], who proposed an explanation for the Sun's energy production from the fusion of light nuclei species, and Hans Bethe [2], in which he gave a more detailed analysis of almost continuous energy release from proton-proton chain reactions, based on Eddington's hypothesis, stand out. Since then, it has been of special interest to replicate similar conditions on Earth and thus to obtain self-sustained fusion reactions.

Most of the current efforts to achieve these conditions have been from the establishment of magnetic confinement devices, where a hot, fully ionized plasma is kept inside the machine with intense magnetic fields. Among the most well-known and successful experimental magnetic confinement devices is the tokamak, a torus-shaped chamber where the combination of closed magnetic fields holds the fuel.

Despite significant advances in the field of tokamak physics, there are still important challenges that need to be addressed. To better optimize fusion output power, it is necessary to increase certain plasma parameters. However, it has been noticed that increasing these parameters beyond operational limits

can cause instabilities in the confined plasma and trigger disruptions, which can seriously damage the integrity of the machine. Thus, disruption prediction is a critical subject for current and future tokamaks, such as the International Thermonuclear Experimental Reactor (ITER), which expects to demonstrate the feasibility of this technology.

The complexity of the physical phenomena involved in the dynamics of disruptions has not yet, however, allowed a complete theoretical understanding and the establishment of real-time disruption prediction systems which can detect all disruptive discharges. With a large number of diagnostics available in the device, researchers are putting effort into the study of the evolution of plasma signals and in the of use advanced algorithms and statistical methods on these data to better understand disruptions.

This work aims to contribute to this field, both on the study of predictive capabilities with the use of a specific diagnostic, as well as in the physical insight from the analysis outcome.

1.1 Tokamak Physics

In tokamak devices, a hot plasma is confined within an axisymmetric torus-shaped vacuum chamber by a superposition of magnetic fields. A set of external toroidal field coils, which carry current in the poloidal direction, creates a toroidal magnetic field B_ϕ. If we take into account the analysis of single-particle motion, one can infer that having only a toroidal field would not be sufficient. A poloidal field B_θ is then generated by the induced plasma current in the toroidal direction through an iron core in the torus center, and external coils, which work as the primary of a transformer and with the contained plasma acting as a secondary winding. To prevent toroidal forces from expanding plasma outwards, a set of vertical field coils are also included [3]. The combined geometry from the established magnetic fields disables major particle drifts and plasma radial expansion, providing a stable configuration.

The magnetic and current field lines, **B** and **J**, follow helical paths along the toroidal flux surfaces, where the field flux is constant. Using the general properties of magnetohydrodynamics (MHD) equilibrium, there is no component from both **B** and **J** perpendicular to these surfaces [3]. The chosen configuration allows radial pressure balance through the combination of magnetic tension and magnetic pressure. A representation of the proposed geometry can be seen in figure 1.1.

The induced toroidal current is a heating source due to the plasma's resistivity, η. However, using the Spitzer resistivity relation [5],

$$\eta = \frac{\pi e^2 m_e^{1/2}}{(4\pi\epsilon_0)^2 (k_B T_e)^{3/2}} \ln\Lambda \tag{1.1}$$

where $\ln\Lambda$ is the Coulomb logarithm, k_B is the Boltzmann constant, ϵ_0 is the vacuum permittivity, and e, m_e and T_e are the electron's charge, mass, and temperature, respectively, it is possible to understand that $\eta \sim T^{-3/2}$, and, for instance, at high temperatures, additional heating methods are required since ohmic heating will be less effective. As auxiliary heating sources, the injection of neutral beams (NBI) [6]

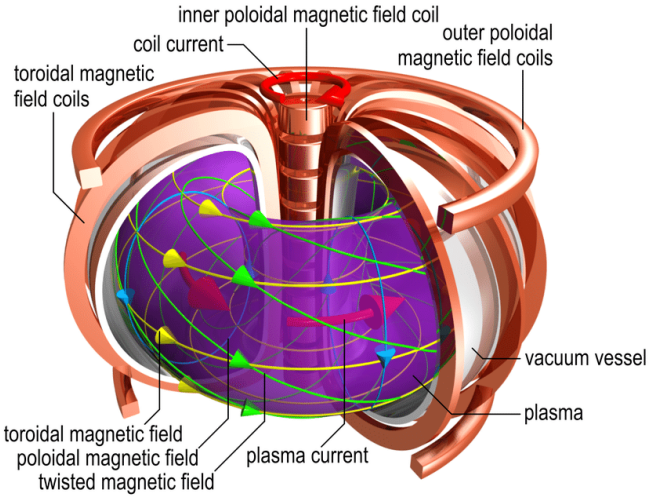


Figure 1.1: Schematic of a tokamak. Source: [4]

and the ion cyclotron resonance heating (ICRH) [7] are some mechanisms currently in use. In the first method, a high energy beam of neutral particles is injected into the plasma, where through Coulomb collisions the energy and momentum are transferred to the background plasma, while in the second the absorption of electromagnetic waves close to the oscillation frequency of the ion species of the plasma also lead to energy transfer and, therefore, in the rise of the plasma temperature.

The stability of the plasma confinement can be characterized by two important relations, the safety factor q and the β factor. As stated, in equilibrium, each magnetic field line follows a helical path around the torus on its magnetic surface. Considering a poloidal plane, the field lines will cross that plane after a toroidal rotation of 2π. If they return to the initial position in that plane after $\Delta\phi$ toroidal crosses, the safety factor can be expressed as $q = \Delta\phi/2\pi$. The safety factor can also be defined as the ratio between the toroidal magnetic flux, $\delta\phi$, and the poloidal magnetic flux, $\delta\theta$. In large aspect ratio tokamaks ($R_0/a \gg 1$, where R_0 is the major radius and a the minor radius of the plasma) equation 1.2 is equivalent [8],

$$q = \frac{rB_\phi}{R_0 B_\theta} \tag{1.2}$$

where r is the radius of the flux surface. If q assumes a rational number m/n, the field line matches its initial position in the poloidal cross-section after m toroidal and n poloidal rotations around the torus axis, with m and n as integer values, and thus expressed with equation 1.3.

$$q = \frac{m}{n} \tag{1.3}$$

Figure 1.3 shows an example of a given q profile, where the magnetic field line returns to its initial position in the magnetic surface after four turns. High values of q contribute to greater stability of the plasma confinement [3], as MHD stability sets a minimum limit to the q values (see Chapter 2).

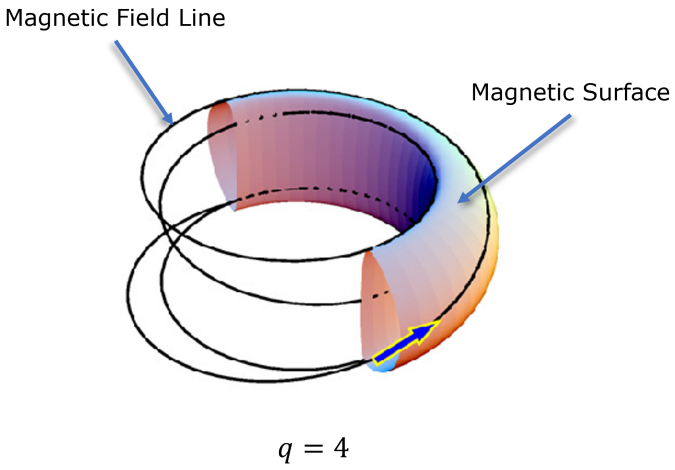


Figure 1.2: Illustration of the q profile in a tokamak, with $q = 4$. Adapted from: [9]

The β parameter measures both the efficiency of confinement of the plasma with the magnetic field and the stability against pressure gradients and can be defined as the ratio of the plasma pressure to the total magnetic pressure, as seen in equation 1.4,

$$\beta = \frac{\langle p \rangle}{B^2/\mu_0} \tag{1.4}$$

where $\langle p \rangle$ is the average plasma pressure and μ_0 is the vacuum permeability. Similarly to q, the achievable β for a stable operation has certain limits [10]. As it will be seen further, increasing or decreasing these parameters to critical values can trigger the growth of instabilities in the plasma, imposing operational constraints on the machine's operation.

1.2 Disruptions

To improve fusion performance, one must work closer to the mentioned operational limits of the tokamak. However, this can trigger the growth of instabilities, which may cause the deterioration of the confinement. This deterioration can become uncontrollable to the point where the confinement is lost, leading to a disruption.

A disruption is a violent and abrupt event where considerable amounts of energy and force loads can be released to the surrounding structures of the plasma, possibly damaging them. Figure 1.3 shows one of the most visible aspects of a disruption, with the deterioration of the interior wall of the vacuum vessel by generated runaway electrons (see Chapter 2). Large parts of the energy stored in the plasma are released in a small fraction of the experiment, typically in a much larger timescale than the Alfvén time τ_A [11]. In major disruptions, this is followed by a complete quench of the plasma current and, for instance, the discharge itself, while in minor disruptions there is no complete current

quench, thus possibly recovering plasma confinement [12]. Apart from eventual causes linked to the human operation, these events are the direct or indirect consequence of MHD instabilities. One of the most known instabilities linked to disruption triggering is the locked mode [13] (see Chapter 2). The disruptions due to the development of locked modes are the main focus of this work.

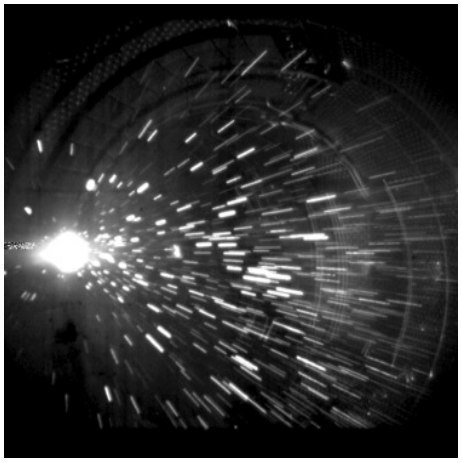


Figure 1.3: Runaway electrons and material sputtering during a disruption at Tore Supra tokamak. Source: [14]

Disruptions impose some limitations in the further development of current and future tokamaks due to their destructive potential and their influence on the operational space of the machine, imposing lower and upper constraints in the plasma confinement parameters, and, for that reason, in maximum achievable output power. This is also a critical issue for the feasibility of projects like ITER [15, 16].

It is, therefore, crucial to develop methods that can either detect the early-stage development of disruptions to prevent them and regain plasma confinement or mitigate its effects on the device with disruption mitigation systems (DMS). The first issue is addressed with disruption prediction techniques.

1.2.1 Disruption Prediction

As stated, tokamak disruptions may be triggered due to the development of instabilities in the plasma. One way to tackle this issue would be with first-principle approaches, in which one can, for example, track these instabilities in real-time with physical-based models, for instance, based on MHD theory. However, some limitations may arise due to the incomplete physical understanding of the disruption mechanisms [17].

Another approach is to rely on the establishment of threshold values for certain plasma parameters, instability amplitudes, and other diagnostic data. However, since these values can fluctuate according to the different experimental scenarios, it is also not completely feasible to use in disruption prediction systems, where there may be, consequently, a considerable number of missed or false disruption alarms [18].

More recently, and also due to the previous limitations, the physics community turned their objective to the use of machine learning algorithms and other advanced statistical methods. This use is also justified by a large number of diagnostics present in some of the experimental tokamaks, with considerably high sampling rates, generating large amounts of data per experiment (a 30 s experiment in the Joint European Torus (JET) tokamak can deliver up to 50 GB of data [19]). These methods have already shown interesting results (see Chapter 3).

The implementation of the mentioned methods is generally based on the use of several diagnostic signals as an input feature vector, defined as $\mathbf{x} = \{x_1, x_2, ..., x_i\}$, with i as the signal index, where they are given to a mapping input function, $y(x_i)$, to separate disruptive and non-disruptive behavior according to present patterns. The way $y(x_i)$ is defined and the type of input mapping depends on the contextualization of the problem in the more suitable machine learning domain. As an example, one can infer the normalized probability for disruption in a given time step of the experiment, by minimizing the difference between a model prediction, $\hat{y}(x_i)$, and a real value $y(x_i)$ with previous offline examples (this is also known as the training process in supervised learning domain). A threshold value in the normalized probability would be used to set an alarm to trigger an actuator or DMS in the device. For a more detailed explanation of these methods see Chapter 3.

Some precautions have to be, however, taken into account when developing these systems for real-time deployment. Cross-machine validation of the developed models is one of them, where researchers try to apply the disruption predictor developed with data from a certain tokamak across other machines. In addition, researchers are targeting performance objectives for future tokamak devices such as ITER, where warning times are critical and the number of missed disruption alarms must be residual.

1.2.2 ITER Requirements

For ITER high-performance operation with deuterium-tritium plasma (DT) it is expected a fusion power gain Q = 10, with a toroidal magnetic field B_ϕ = 5.3 T and a plasma current I_p = 15 MA [20]. Having into account these values, guidelines for disruption prediction methods were delivered to achieve operational requirements. These requirements are especially focused on the trigger of the DMS in ITER. The most relevant can be set in the minimum time window for the predictors to give an alarm and their capability to correctly identify disruptions.

Based on the current DMS technologies, such as shattered injected pellets (SPI) or massive gas injection (MGI), as well as their response time in current operational tokamaks, a minimum warning time of 30 ms is generally accepted to successfully predict a disruptive event [21] in advance. In machine learning standards, this is also known as a limit to define a true positive discharge (disruptive and $t_{alarm} > 30$ ms) or a false positive discharge (disruptive and $t_{alarm} < 30$ ms) experiment. Also, the success of disruption prediction based on ITER requirements is measured in terms of the percentage

of correctly predicted disruptive discharges. The acceptable fraction of correct predicted disruptive discharges (or accuracy) is between 95% to 99%. In contrast, missed or later triggers (also denominated as false negative rate) should be less or equal to 5%. Studies based on machine learning algorithms aim, therefore, to achieve these requirements so that integration in real-time systems is possible.

1.3 Goals

The main goal of this work is to evaluate how MHD activity can contribute to the disruption prediction field, by training a neural network model which receives a spectrogram sample coming from a magnetic field coil installed at JET tokamak and analyzing its prediction performance results. Also, we try to obtain the most important frequency bands of the spectrogram that the model retrieves to deliver a given output.

In the first step, the locked mode signal is studied using binary classification metrics, in order to define a value for which a given discharge is considered disruptive. The chosen magnetic coil data is then processed into frequency space for each experimental pulse and given as an input to the implemented model, which is trained to predict locked mode disruptions.

In the next stage, a convolutional neural network (CNN) model is set to solve a binary classification problem, receiving a spectrogram sample of a given pulse to determine if it will disrupt due to a locked mode. To interpret the model's inference, a technique for discriminative localization is applied. This particular stage is important because it allows to analyze if the features retrieved by the model converge with the physical interpretation, namely with possible MHD precursors present before the disruption. It also opens the possibility for researchers that use data-driven methods in disruption prediction, with other sources of 0D, 1D, or 2D input data, to analyze the results with interpretable machine learning frameworks.

1.4 Thesis Outline

The structure of this work can be summarized as follows:

- Chapter 2 introduces the theoretical concepts and evolutionary characterization of disruptions, as well as a brief description of the MHD instability development and relevant activity involved.
- Chapter 3 presents the state-of-the-art methods to deal with disruptions (avoidance and mitigation techniques), including achieved milestones and obstacles. The main focus is on methodologies based on predictive algorithms and their ability to discriminate disruptive and non-disruptive behavior.

- In chapter 4, the experimental setup of this work is introduced, with an overview of the available diagnostics system in the device, as well as the chosen instruments for probing magnetic oscillations and the locked mode signal. The treatment of data coming from magnetic oscillations for representation in time and frequency is mentioned at the end of the chapter.
- Chapter 5 describes the proposed methodology for using spectrograms from the mentioned experimental setup to tackle the disruption prediction target and shows how the amplitude of locked mode is considered in disruption definition. It includes the deep learning algorithm implementation and a description of the technique used for the model interpretability.
- The results of model training and its prediction performance metrics are given in chapter 6. Additionally, the physical insights are studied by analyzing the CAM technique and possible links to relevant MHD activity that can contribute to disruptive behavior.
- To conclude, chapter 7 summarizes the main contributions from the proposed approaches and possibly additional work for future research.

2

Background

Contents

A background basis on the physics of disruptions must first be given to better understand the processes addressed in this work. In section 2.1, we will first look for the chain of events that typically characterizes a tokamak disruption, namely through the temporal evolution of some of the machine's diagnostics and plasma parameters. The causes for these events begin to be discussed first through the study of the operational space of a tokamak in section 2.2.

The contextualization on the use of MHD theory in disruptions is presented in chapter 2.3, which includes a brief theoretical summary on the development of instabilities and a more careful description of those that can more intensely lead to plasma disruptive behavior.

2.1 Events leading to a disruption

Most tokamak experiments have a large number of diagnostic systems that allow monitoring on both plasma and machine state parameters such as density, current, electron cyclotron emission (ECE), among others. The evolution of these parameters can be substantially different in cases where plasma confinement is lost, allowing a temporal characterization of the disruption. The first time-trace chart

of figure 2.1 shows the difference between non-disruptive and disruptive discharges in JET, where the evolution of the plasma current at the plateau region, for example, is different [22].

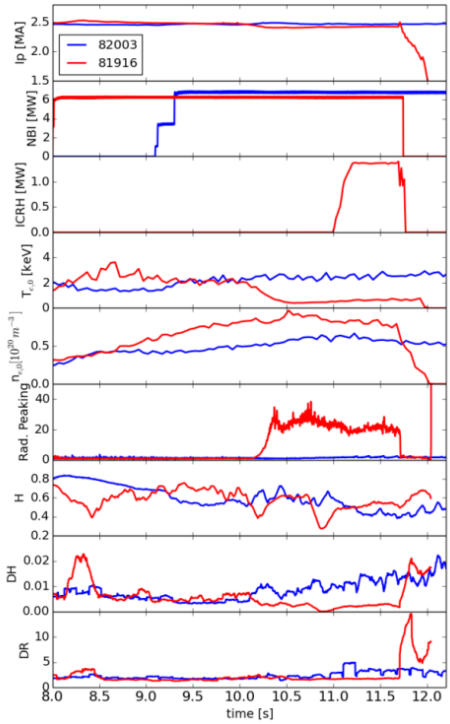


Figure 2.1: Time variation of different plasma diagnostic signals in disruptive (red) and non-disruptive (blue) discharges. Source: [22]

The processes involved can be summarized into different stages. First, some of the plasma parameters may show significant changes in operating conditions which can be generally observed (such as the electron's temperature T_e or magnetic oscillations), leading to the onset of MHD instabilities when higher stability limits are met. The described regime is known as the precursor phase. Its duration depends on the growth rate of the instabilities, which can last several milliseconds before thermal energy dissipation. Among the most common precursors is the locking of magnetic islands which do not rotate relative to the laboratory frame (mode-locking) [23].

Then, if the instability surpasses a critical amplitude threshold, thermal energy loads are released to the first wall of the device and the plasma column cools down, with a decline in the temperature profile to the 10 eV regime [24] and a maximum peak on plasma current after the flat-top section. A negative spike in loop voltage is also observed, which is related with the toroidal electric field around the vessel, and associated with the flattening of the plasma current [24] (see figure 2.2). These events constitute the thermal quench (also known as energy quench). It is also a considerable fast phase of a disruption, usually lasting no longer than 1 ms.

The disruption culminates in the current quench phase, with a rapid decay in the plasma current

profile. The torus vacuum vessel is subjected to considerable electromagnetic force loads produced by $\mathbf{j} \times \mathbf{B}$ and the plasma column can be vertically displaced. This vertical displacement can occur even before the thermal quench [11].

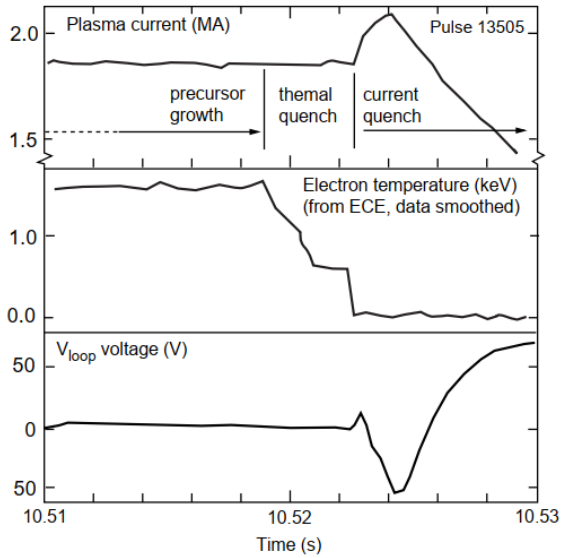


Figure 2.2: Example of disruption development at JET. Source: [24]

It is also commonly observed during the thermal and current quench phases the generation of high energy relativistic electrons (runaway electrons) in the MeV energy scale, which can hit and transfer their energy into the vessel walls, causing material damage. The physical process for their appearance is related to the establishment of an electric field with a parallel component to the magnetic field by inducing a toroidal current in the plasma, which exerts an accelerating force on plasma electron species. Although the Coulomb collisions may slow the movement of these particles, if we consider the electron's collision frequency ν_{ee}, and given approximately by equation 2.1 [3],

$$\nu_{ee} \approx \frac{e^4 n_e \ln \Lambda}{2\pi e_0^2 m_e^2} \frac{1}{v_e^3 + 1.3 v_{Te}^3} \tag{2.1}$$

where v_e and v_{Te} are the electron's kinetic and thermal velocity, respectively, and n_e is the electron's density, shows that as the electron reaches a velocity comparable to the speed of light c, the effect of collisions to counter the electric field acceleration becomes negligible. The required field to trigger runaway electrons in plasmas is given by equation 2.2,

$$E_c = \frac{n_e e^3 \ln \Lambda}{4\pi e_0^2 m_e c^2} \tag{2.2}$$

which is also known as the Dreicer field [25, 26]. Conditions during the regular functioning of the experiment prevent the electric field to be above the Dreicer limit, thus almost no runaway electron is

generated. However, as we lower the density, usually during the thermal quench, the value of the critical electric field is considerably reduced, triggering the release of these electrons whose velocity is typically much higher than the thermal velocity v_{Te}. At first, just a small population from the Maxwellian distribution tail will have a critical velocity, which due to diffusion, is the primary mechanism. With enough seed population, primary electrons can excite thermal electrons and form an avalanche of secondary electrons, leading to exponential growth. This is known as the secondary mechanism. Due to the impact of these electrons in the vacuum vessel of the machine and its destructive potential, these processes must also be taken into account and are part of the design constraints of future tokamaks [27, 28].

2.2 Operational Limits

The performance of a tokamak is limited by the machine's operational space, namely through the safety factor q, the β parameter, and the plasma electron's density n_e. Surpassing these limits can cause either the cooling of the plasma column due to radiation or the trigger of MHD instabilities, then causing the degradation of confinement and a disruption.

The maximum achievable density is limited by ECE, Bremsstrahlung, and other sources of radiation due to the accumulation of impurities. When radiation becomes dominant in heat hollow, the edge region of the plasma column cools down. According to equation 2.3, the density limit can be scaled as a function of the total heating power P_h [29],

$$n_e^{lim} \propto \left(\frac{P_h}{Z-1}\right)^{1/2} \tag{2.3}$$

where Z is the atomic number of the plasma impurities. This suggests that improving external heating or reducing impurities would increase the achievable limit. However, the deduced value is constrained by other mechanisms, such as particle transport at the edge of the plasma column. Murakami et al. [30] have studied density limits in ohmic plasmas from the analysis of different tokamaks and it was observed that the data practically fitted to a dependence of the line averaged density $\bar{n}$ on B/R_0. Later on, M. Greenwald et al. [31] studied the data from a set of different experimental tokamaks and reached an empirical relation for $\bar{n}$ (in $10^{20}\mathrm{m}^{-3}$), which can be expressed as a function of the elongation parameter for the plasma poloidal cross-section, κ, as seen in equation 2.4,

$$\bar{n} = \kappa \bar{J} \tag{2.4}$$

where $\bar{J}$ is the inline average plasma current density. Figure 2.3 shows the Greenwald limit as function of the experimental density range for three different tokamaks, where the normalized Greenwald limit ($n/\bar{n}$ or n/n_G) is mostly less or equal to 1. Interconnection between the density limit and other tokamak

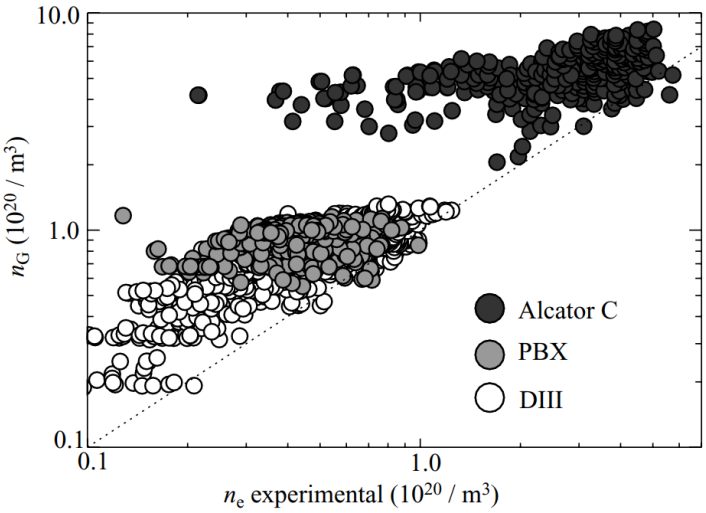


Figure 2.3: Observation of the Greenwald limit for different tokamaks. Source: [3]

parameters, such as the safety factor q, is relevant because as the plasma contracts, current profiles are adjusted, thus also changing the value of q.

Constraints in the safety factor are usually related to low-value limits. It is considered for a stable operation in terms of the safety factor at the edge $q_a > 2$, as instability analysis provided by MHD theory suggests that as the magnetic surface moves towards the plasma edge due to an increase in the plasma current, $m = 2$ and $n = 1$ kink modes [32] are likely to destabilize the plasma column. It is also related to the maximum achievable density through the analysis of the Hugill diagram, which plots the inverse of the safety factor as function of the Murakami parameter. From the analysis of figure 2.4, the majority of the data is below the inverse edge safety factor of 0.5. Additionally, the external heating power can considerably increase the achievable density, as already discussed. The stable and unstable spaces are well discriminated with the possible onset of MHD disruptive instabilities such as external kink modes and tearing modes, with $q < 2$.

Moreover, with the β parameter, high performance is achieved at maximum pressure for a given toroidal field. The ideal β parameter against MHD instabilities was deduced by Troyon et al. [33]. Using optimization techniques for analyzing experimental setups with different current and pressure profiles, the calculations suggested that a non-circular, elongated cross-section would improve significantly the achievable β limit, resulting in an empirical relation defined as

$$\beta = 0.14\frac{\epsilon\kappa}{q} = \beta_N \frac{I}{aB} \tag{2.5}$$

where ϵ is the aspect ratio, defined as $\epsilon = a/R_0$, and β_N is a numerically deduced coefficient. However, elongated plasmas are more susceptible to downward and upward movements towards the vacuum

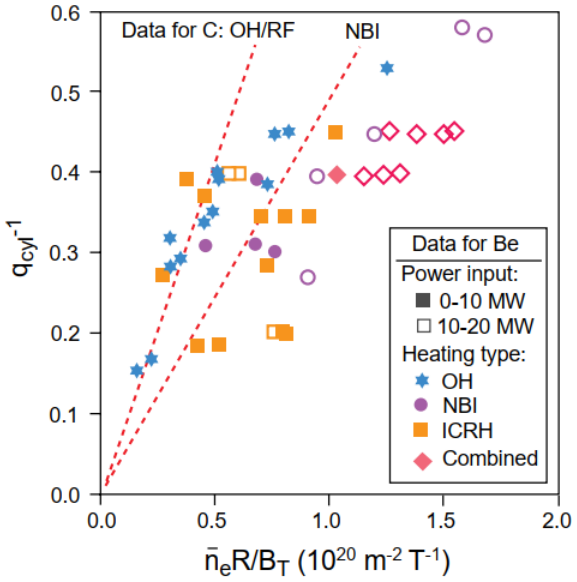


Figure 2.4: Hugill diagram with JET experimental data. Source: [10]

vessel and, thus, to vertical displacement instabilities [3]. High-pressure gradients can also induce other instabilities, such as ballooning modes [34]. The achievable β is also limited by resistive wall modes (RWM), which occur due to the presence of finite resistivity and conductivity in the tokamak wall that surrounds the plasma [35].

2.3 Magnetohydrodynamics Model

In order to describe the plasma macroscopic behavior in a tokamak, a more rigorous model than single-particle motion in terms of self-consistency characterization (i.e. influence of each particle on the generated electromagnetic fields [3]) is necessary, which takes into account the contribution of both electrons and ions, thus having a simplified single-fluid model to analyze its equilibrium and stability. The MHD model provides a framework that corresponds to these requirements.

Assuming an ideal fluid with no dissipation effects, the dynamic parameters can be determined by a set of fluid equations and the Maxwell's equations. The first is the continuity equation,

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{2.6}$$

where ρ is the plasma density and $\mathbf{v}$ is the fluid velocity. Considering the force equations, one reaches the momentum equation,

$$\rho(\partial_t \mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v}) = \mathbf{j} \times \mathbf{B} - \nabla p \tag{2.7}$$

where p is the plasma pressure, **j** the current density, and **B** the magnetic field. To determine both p and ρ, the adiabatic equation is used,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{p}{\rho^{\gamma}}\right) = 0 \tag{2.8}$$

where γ is the adiabatic coefficient, with a numerical value equal to γ = 5/3 (assuming an ideal monoatomic gas, with 3 degrees of freedom). The ideal Ohm's law is written as

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = 0 \tag{2.9}$$

with **E** as the electric field. Finally, the Maxwell's equations (namely the Ampère's law, the Faraday's law, and the magnetic divergence equation) are used to close the system.

$$\nabla \times \mathbf{E} = -\partial_t \mathbf{B} \tag{2.10}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} \tag{2.11}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{2.12}$$

These are the ideal MHD equations. In some situations, it is necessary to take into account dissipation effects and thus the resistivity term, η. Adding the dissipation term, equation 2.9 is modified and given by equation 2.13.

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = \eta \mathbf{j} \tag{2.13}$$

After linearization of the previous system of equations for stability analysis, assuming the perturbed quantity amplitudes to be much smaller than the equilibrium state, a displacement perturbation vector, ξ, defined in space and time, and given by equation 2.14, is introduced into the linearized MHD equations.

$$\xi(\mathbf{r}, t) = \xi(\mathbf{r}) e^{i(m\theta - n\phi + wt)} \tag{2.14}$$

In equation 2.14, θ and ϕ are the poloidal and toroidal angles, respectively, m and n the poloidal and toroidal mode numbers, and w is the frequency of the perturbation. The ξ term is introduced within the perturbed fluid velocity term, v_1, as seen in equation 2.15.

$$v_1 = \frac{\mathrm{d}\xi}{\mathrm{d}t} \tag{2.15}$$

In non-ideal MHD, the frequency w has both real and imaginary parts (w_r and w_i), and the stability of the mode is determined by the sign of w_i. If w_i is lower than 0, then the system is said to be unstable, and

stable otherwise. By solving the linearized equations in terms of ξ, one finds the eigenvalue equation,

$$-w^2\rho\xi = F(\xi) \tag{2.16}$$

where F is called the force operator. It is by isolating F that one finds the stability limits for a given displacement. To conclude this analysis, the boundary conditions must be specified. A more detailed explanation of the MHD stability analysis can be seen in reference [3].

2.3.1 Tearing Modes

When taking into consideration the resistive MHD theory (adding the dissipation term in Ohm's law) a specific type of unstable modes driven by current gradients, called tearing modes, can appear. Destabilizing these modes can produce poloidal regions where there is a reconnection of the magnetic field lines, reflecting a change of the magnetic topology and violating the frozen-flux condition due to the resistivity (for ideal MHD, the magnetic flux through a surface moving with the plasma is constant [8, 36]). This is typically associated with the formation of "magnetic islands" (see figure 2.5), where the magnetic field lines break and reconnect, forming island-like structures.

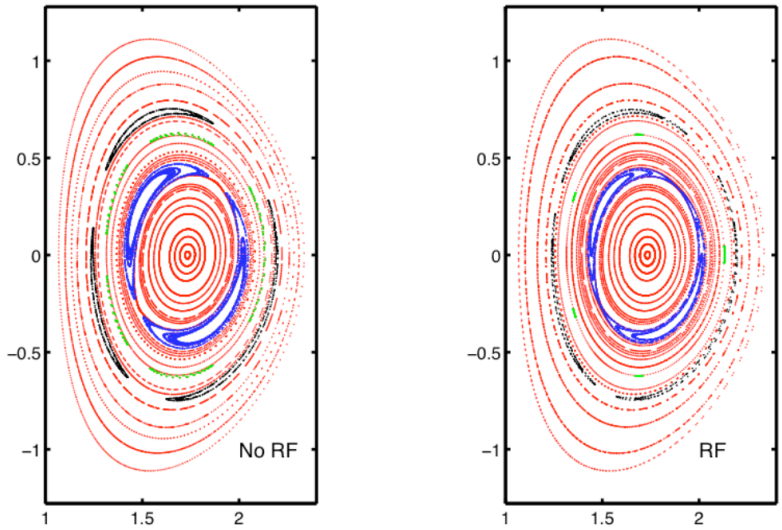


Figure 2.5: Evolution of interior magnetic islands in a poloidal cross-section view. Source: [37]

The stability of the tearing mode can be determined by the "tearing parameter", introduced by Furth et al. [38], and defined as,

$$\Delta' = \frac{1}{B_r}\frac{\partial B_r}{\partial r}\Big|_{r_s+w_s/2}^{r_s-w_s/2} \tag{2.17}$$

where B_r is the radial component of the magnetic field, r_s is the resonant magnetic surface and w_s the magnetic island width. If $\Delta' > 0$ then the mode is said to be unstable. The width evolution of the formed

island can be approximately written as function of the tearing parameter [8].

$$\frac{\mathrm{d}w}{\mathrm{d}t} \approx \frac{\eta}{2\mu_0}\Delta^{'} \tag{2.18}$$

Also for a typical deuterium plasma, the growth rate is expressed as a function of the resistive diffusion time τ_R and Alfvén transient time τ_A.

$$\lambda^{-1} \approx \tau_R^{3/5}\tau_A^{2/5} \tag{2.19}$$

In [39], a degradation on the confinement time $\tau_E = \Delta W/\Delta P_i$, where W is the stored energy in the plasma and P_i the power input, was predicted for tearing modes with $m = 2$, $n = 1$. A β limit reduction has also been observed with tearing modes with $m = \{3, 2\}$ and $n = \{2, 1\}$ in tokamak experiments [40]. Even though $\Delta^{'} < 0$ in the previous case, the plasma configuration gets limited, with a formation of a seed, low-scale magnetic island. This was an example of the occurrence of a neoclassical tearing mode (NTM) [41], a similar instability to a classical tearing mode mainly caused by perturbations in the bootstrap current (current generated due to the colision of trapped particles in poloidal orbits and the particles passing along the field lines, as well as due to density gradients), facilitating the growth of the magnetic island. As opposed to the tearing mode, NTMs need an initial seed island with sufficient width to destabilize. NTMs can particularly contribute to confinement degradation, in particular with unstable $m = 2$, $n = 1$ modes. Figure 2.6 shows a spectrogram observation of a NTM with $n = 2$ that starts at around $t = 6.2$ s, with a frequency of approximately 17 kHz.

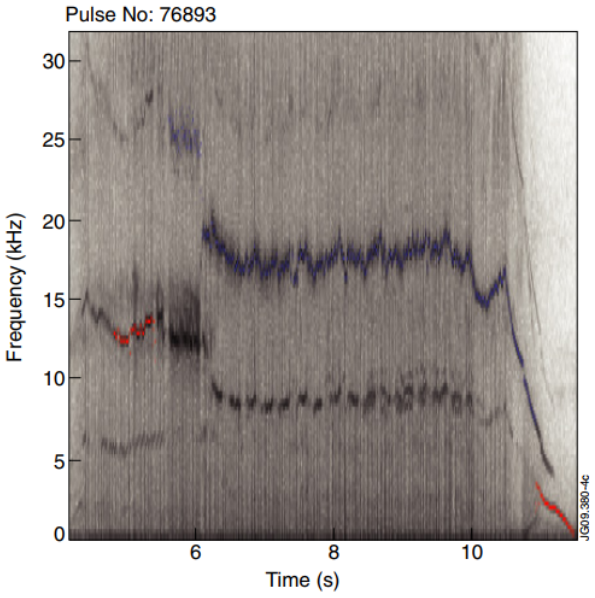


Figure 2.6: Spectrogram of a pick-up coil data at JET with NTM behaviour from $t = 6.2$ s. Source: [42]

2.3.2 Mode Locking

During the growth of a tearing mode or a NTM, the width of a magnetic island can be large enough to have considerable interaction with the tokamak's first wall. This interaction can cause the deceleration of the mode, eventually stopping its rotation and making it to lock. This behavior is visible through the nulling of the frequency component of a magnetic signal, and the rise on its amplitude, as seen in figure 2.7. If the magnetic field coils of a tokamak are misaligned they generally produce error fields, which are one of the common excitation sources of locked modes. The deceleration is attributed to applied torques in the vacuum vessel wall [43]. In some cases, the mode can lock before any rise in amplitude, as experimentally observed in a study by Sweeney et al. [44] with $m = 2$, $n = 1$ NTM modes. An additional mechanism allows their development due to a perturbation field produced by the bootstrap current [23].

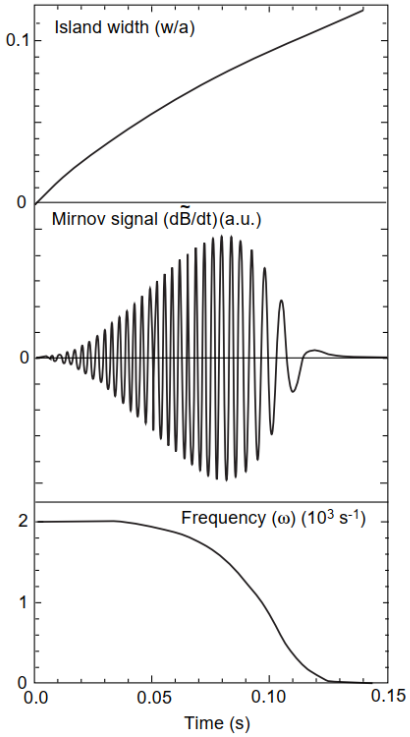


Figure 2.7: Simulation of mode-locking with a JET plasma. Source: [45]

Locked modes are already well known to have an intrinsic relationship with tokamak disruptions, such as in JET [13, 46], and are also a target for ITER development [47]. As it will be seen in chapter 3, advances in disruption prediction and mitigation systems rely heavily on mode-locking as part of their input data. Although locked modes are a crucial factor in disruption development, they may not allow a complete fast mitigation action since they can happen very close to the disruption instant in some cases.

2.4 Conclusion

In this chapter, the mentioned MHD activity was mainly focused on the onset of tearing modes and NTMs that slow down their rotation when interacting with the machine walls, eventually locking. This is a typical behavior observed in tokamak disruptions.

The use of the locked mode signal alone has some limitations. One can ask what mechanisms could act as possible MHD precursors before mode-locking itself. Having this information could increase considerably the time window in which the mitigation systems of the machine could act. Combining both the frequency information about the magnetic field oscillations and the amplitude of the mode, this work expects to give some hints on that matter.

3

Related Work

Contents

In this chapter, we will explore some of the work done in the field of tokamak disruptions in order to understand the possible framework and capabilities of the proposed approach. We first look at some examples of mitigation systems present in tokamaks. Moreover, a revision is done in the use of data-driven approaches to disruption prediction. We also go through on what has been done in terms of insight or knowledge retrieval with these approaches, that is to say, how one can extract the decision rules from machine learning models, as well as its deduced features to make a certain prediction.

3.1 Disruption Mitigation

When handling disruptions, two main approaches are considered. On one hand, one wants to avoid the plasma to surpass the operational limits of the machine and thus to prevent disruptions from happening. Prevention techniques are needed to achieve this goal, which may include the detection of disruptive precursors, for example. On the other hand, if that is not possible, then one needs to minimize any force loads on the vessel walls and material wear. In this scenario, the mitigation techniques are used to avoid any major damage to the device.

In JET, a disruption mitigation valve (DMV) based on the injection of high-pressure noble gases to

the plasma was tested [48]. This injection enables a fast discharge termination, radiating most of plasma energy stored. To trigger these systems, mainly two signals, including the locked mode amplitude, are used.

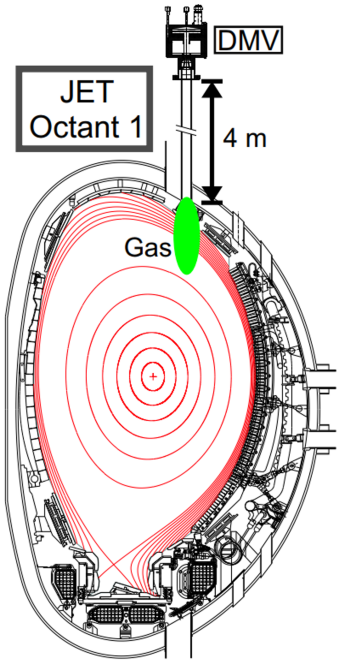


Figure 3.1: Cross-section view of JET with the DMV system. Source: [49]

The trigger can be activated if the locked mode amplitude surpasses a threshold at a given instant. The defined thresholds, however, may depend on the analyzed dataset and have some fluctuations on optimal values [50], thus some disruptions could be ignored by the system. Some pulses could also be falsely identified as disruptions.

The injection of impurities is also used with a SPI installment to regulate the plasma-material interactions [51]. In ITER standards, using the mentioned system should reduce the total stored plasma energy by at least 10% of its initial value.

3.2 Machine Learning on disruption prediction

Despite the efforts to deliver a complete theoretical framework for disruptions [11], the considerably high percentage of failed disruption alarms by only setting a threshold on certain plasma parameters and the challenges faced when implementing first-principle approaches have led to the use of machine learning algorithms which can read the diagnostics data and build non-linear relations between them to detect disruptions. Mainly two types of problems are defined when using these methods for disruption prediction. One can define a probability of a given discharge to disrupt or calculate the time to a disruption

instant (TTD). This work focuses on the first approach. Also, when dealing with these algorithms, different types of methods arise (mainly supervised and unsupervised methods), depending on the defined task.

3.2.1 Supervised Learning Methods

With supervised learning predictors, a label is priorly given to a training sample. That is, each discharge is classified as disruptive or non-disruptive, where input and output are both fed to the algorithm. This type of training is one of the most used approaches, that can go from the use of more classical machine learning algorithms, such as random forests [52], support-vector machines (SVM) [53, 54], gradient boosted trees [55], or deep learning algorithms [56]. At JET, a real-time predictor [57] (APODIS) based on SVM that uses 7 input signals was tested. It shows a significant percentage of correctly predicted disruptions (93%) and an improvement in the time window for correct predictions.

Additionally, the use of deep learning models, which can go from the use of multi-layer perceptrons [56] to more complex systems such as Convolutional Neural Networks (CNN) [58], Recurrent Neural Networks (RNN) [59], and Long Short-Term Memory (LSTM) networks [60] has shown promising results. Kates-Harbeck et al. [17] proposed a new architecture, called Fusion Recurrent Neural Network (FRNN), which combines the functionalities of both RNN, LSTM, and CNN models. An example of the predictor can be seen in figure 3.2. The FRNN outperforms a classical machine learning method providing a threshold alarm long before the 30 ms mark. The FRNN was trained on JET and DIII-D data using an input vector of more than 15 signals, including 1D and 2D data, reaching more than 95% of the area under curve metric (AUC) when tested on JET data.

3.2.2 Unsupervised Learning Methods

Contrary to the methods described above, no label is given to the training data with unsupervised learning methods. This time, instead of training the system to know what is a non-disruptive or disruptive experiment, unsupervised learning algorithms automatically draw a separation region for these classes without prior knowledge of the output.

In [61] a k-nearest neighbor algorithm was trained on JET data to automatically classify different kinds of disruptions, yielding more than 90% of corrected classifications. Another recent study [62] tests the implementation of a linear equation, based on the computation of centroids to separate disruptive and non-disruptive behaviors. It only uses the locked mode amplitude normalized to the plasma current as unique parameter, and shows interesting results on test data in both the disruption detection rate (with 90% of valid alarms) and in the TTD metric. Also, a clustering technique was tested using only the locked mode signal [63], which presented some improvements when compared with the threshold

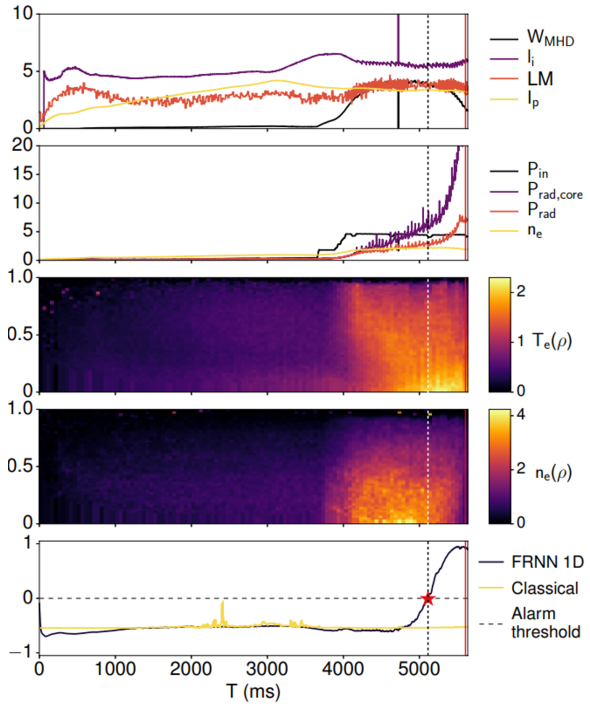


Figure 3.2: Example of a classic machine learning model and deep neural network implementation with scalar 0D signals and 1D profile data over a radial coordinate for disruption prediction on JET. Source: [17]

definition approach and the APODIS system.

Another goal when using unsupervised learning methods is to detect possible outliers within the diagnostics data in the normal functioning of the machine, which could correspond to a precursor development phase of a disruption (anomaly detection). This was done, for example, in a recent study by Ferreira et al. [64]. A variational autoencoder replicates the radiation profile of JET data to detect anomalies during the precursor phase of a discharge. Figure 3.3 shows an example of the application, in which the anomaly score begins to increase when an intense radiation at the plasma core appears, being able to capture the development of a disruption precursor.

Despite the capabilities of unsupervised learning methods, they are beyond the scope of this work, as we needed to provide to a given model an example of both disruptive and non-disruptive examples in advance.

3.2.3 Interpretability

The equations derived from machine learning methods can indeed improve significantly the capabilities to discriminate non-disruptive and disruptive behavior. However, there are some disadvantages if one tries to obtain physical insight due to the complexity of its equations [62]. This insight can be useful to understand better the possible mechanisms that trigger a disruption and the features retrieved by the

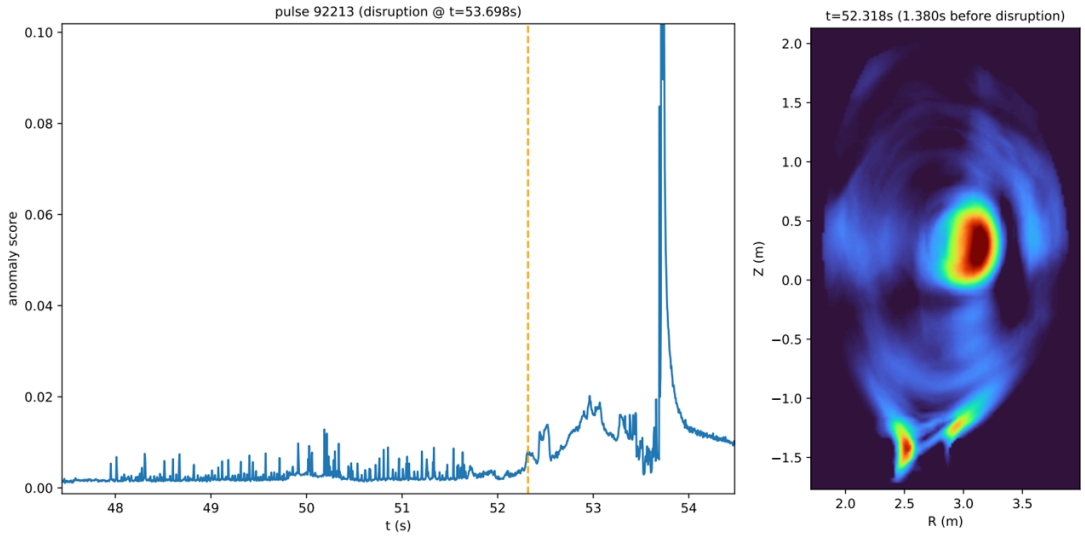


Figure 3.3: Anomaly detection with tomographic reconstruction for precursor analysis. Source: [64]

models. This is even more relevant in deep learning models, often called ”black boxes” due to their multilayer structure. Churchill et al. [58] pointed out the necessity of having interpretability present when machine learning methods such as neural networks, in particular, are applied in disruption prediction, to allow cross-machine validation in transfer learning or to identify particular areas of interest that can be related to the precursor phase, for example. Research regarding this topic is thus also starting to be carried out.

In [65], an equation of the locked mode amplitude that separates non-disruptive and disruptive regions in the operational space was obtained using symbolic regression. This method was able to achieve more than 92% disruptive discharges on a JET data set and the results were comparable with previous machine learning methods. More recently, Rea et al. [66] developed a feature engineering method to improve cross-machine validation of methods and their physical interpretability.

In addition, both explainability and interpretability methods in deep learning have been the subject of research [67]. One of the techniques developed for CNN [68], important in the recognition of digital images and also used in this work, has been already applied in nuclear fusion domain [69] with visible imaging data from the vacuum vessel of a tokamak. Although the developed system was quite capable of classifying disruptive frames, it could not be directly used as a predictor of disruptions.

3.3 Use of MHD activity for disruption prediction

It is well known that locked modes contribute as important disruption precursors. De Vries et al. [13] showed that the locking of NTMs was found as the main source of disruptions in a set of studied disruptive discharges of JET. This kind of MHD activity is then often used as a feature in forecasting methods.

In reference [65] a scaling law for the critical locked mode amplitude was achieved, which depends on other plasma parameters such as the internal inductance, li, producing results that are comparable with other machine learning methods. Also, in DIII-D tokamak data, a recent study [44] shows that the disruptive phenomenon of initial rotating tearing modes could be described by the inductance normalized to the safety factor and the distance between the island and the plasma separatrix, where at least 7% and 4% of disruptions were missed, respectively, when testing the method on new data. Sias et al. [70] developed a disruption indicator for both JET and AUG tokamaks which uses a processing algorithm to the locked mode signal and improves the analysis previously done using only raw data.

Also within the machine learning framework, the locked mode signal was used as the unique feature in an anomaly detection predictor called SPAD [71], validated and tested with JET data. In comparison with the APODIS system, it shows some improvements in disruption detection ($>$ 83% of correctly detected disruptions). It was also shown that a predictor based only on threshold definition delivers a significantly low value of the detected disruption rate and very close to random probability.

Of special interest in the present work is the use of magnetic diagnostics from the tokamak and the analysis of the underlying MHD activity, using it as an input to a deep learning framework. This is not the first time that such kind of input is considered. In [72], measurements from Mirnov coils were used to identify disruptions caused by a $m = 2$ mode precursor using neural networks. It was also stated in [18] that the combination of time and frequency information of MHD activity is a possible improvement technique to apply in reliable predictors. More recently, in Bustos et al. [73], a deep learning model was used to identify, via segmentation, the active oscillation modes from Alfvén waves in MHD spectrograms.

3.4 Conclusion

Despite the capabilities displayed by all the implementations described above, some of them using only the locked mode signal or incorporate it in a set of features, no existent solution has been able whether to predict in real-time or in offline testing the full stack of disruptive events. Therefore, the requirements of ITER are not completely satisfied yet.

Our work is, to the best of our knowledge, one of the first attempts to use interpretable machine learning techniques to the Fourier analysis of magnetic diagnostics data for disruption prediction. More specifically, one wants to use interpretability to identify the MHD behavior that the deep learning model finds more important when predicting a disruption distinguished by the presence of locked modes.

4

Experimental Setup

Contents

In the following chapter, an overview of the JET tokamak device is presented, as well as the specific diagnostic signals that follow the requirements for the proposed approach. These signals include data concerning the locked mode and the measurements of the local magnetic field fluctuations. Furthermore, the used data processing techniques are mentioned and briefly described.

4.1 JET Tokamak

The JET tokamak is an experimental device, located at the Culham Center for Fusion Energy, in the United Kingdom, and part of the European Fusion Programme (EUROfusion). It was established in 1983, reaching its first plasma in that same year.

The toroidal field is generated by a set of 32 coils with a D-shaped cross-section, which surrounds the main vacuum vessel. The vacuum vessel is divided into 8 different octants by the central iron core support limbs. An NBI external source injects energetic beams in the same direction as the plasma current, which can deliver up to 34 MW of plasma heating [74]. Additionally, the ICRH system can also be used as an external heating source by sending electromagnetic waves close to frequencies in the ion cyclotron motion range. JET's main operational parameters are shown in table 4.1, and a more detailed

explanation of the machine can be seen in [75]. JET currently holds an important role as a benchmark

Parameter	Value
Major radius (R_0)	$\sim$ 3 m
Minor radius (a)	$\sim$ 2 m
Maximum plasma current	$\sim$ 5 MA
Maximum toroidal magnetic field	$\sim$ 3.5 T
Pulse duration	$\sim$ 30 s
Maximum elongation (κ)	$\sim$ 1.8
Maximum plasma heating	$\sim$ 50 MW
Input power	$\sim$ 500 MW

Table 4.1: JET operational parameters.

for future tokamaks. It has been operating since 2011 with materials (beryllium and tungsten) similar to the plasma-facing components of the upcoming ITER device (JET-ILW campaigns) [21], allowing a more systematic characterization of the operational performance with ITER setup. It is also the only known machine to achieve the highest experimental value of the fusion energy gain factor with D-T reactions, in 1997 [76], although a higher value of the extrapolated gain factor, Q_{DT}^{eq}, was achieved in JT-60 tokamak [77].

JET's diagnostic system consists of approximately 100 instruments, most of which are divided into several channels to allow measurements along the plasma section. They can measure global properties and parameters from the heating plasma, either from particle-particle or particle-wave interactions. Figure 4.1 shows a schematic of the system of diagnostics installed in JET. In the context of this thesis,

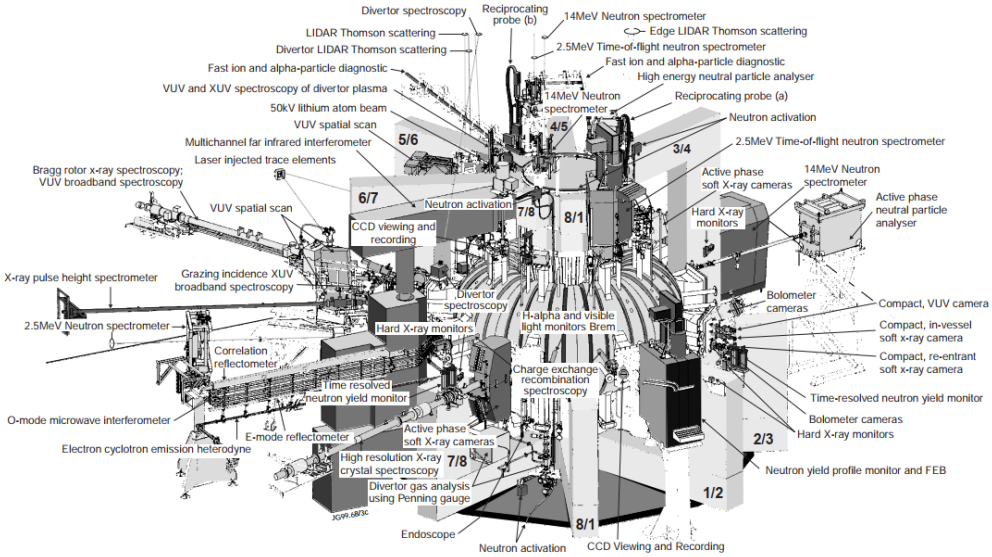


Figure 4.1: Diagnostics installment in JET tokamak. Source: [78]

the most important are the magnetic diagnostics. They consist of a system of coils and probes installed

along the vacuum vessel that measure the magnetic field fluxes, allowing important parameters such as the plasma position on the vessel, the rotation frequency, and plasma current, to be retrieved [79]. In the following sections, the instruments of interest will be described.

4.2 Magnetic Diagnostics

In order to study the plasma equilibrium, stability, and spectral response, a set of inductive, cylindrical coils are used, designated as Mirnov coils. They consist of N loops of titanium wire that measure the time variation of the local poloidal magnetic field, B_θ. Following Faraday's law, a voltage is induced at the wires of the coil due to the variation of the magnetic flux density. The measurement from these coils can be summarized by equation 4.1, and thus by the integral of the product of its effective area A and the time variation of B_θ,

$$\Phi = -N \frac{\mathrm{d}}{\mathrm{d}t} \int B_\theta dA \tag{4.1}$$

where N is the number of turns in the probe.

These coils allow the determination of the amplitude and frequency of the perturbations and, for instance, to study the MHD activity. By combining them at different locations in the vacuum vessel, it is possible to retrieve information about the toroidal and poloidal numbers for a given MHD mode.

4.2.1 H305 coil

A single Mirnov coil was used for this work. It is part of an array of magnetic coils (High Resolution Array Coils) distributed at different toroidal angles and at the same poloidal plane, in the third octant of the vacuum vessel. These coils are specially designated for high mode analysis [80]. Their toroidal location is given by table 4.2 and figure 4.2, where the H305 coil is chosen.

The integrated data acquisition system of the H305 coil provides a maximum sampling rate f_s of 2 MHz, allowing the study of MHD modes up to 1 MHz according to the Nyquist frequency and, for instance, equation 4.2. However, as the frequency range of interest for MHD mode analysis was considerably lower than f_{min}, a downsampling to 125 kHz was made, which is also comparable to other downsample approaches with other diagnostics [58].

Coil	ϕ **(º)**
H301	-13.00
H302	2.94
H303	13.11
H304	18.74
H305	20.38

Table 4.2: Toroidal location of the different Mirnov coils at JET tokamak.

$$f_{min} = \frac{f_s}{2} \tag{4.2}$$

The raw data of the H305, as it will be seen further, is later processed in a time and frequency representation through Fourier analysis in order to study the MHD activity.

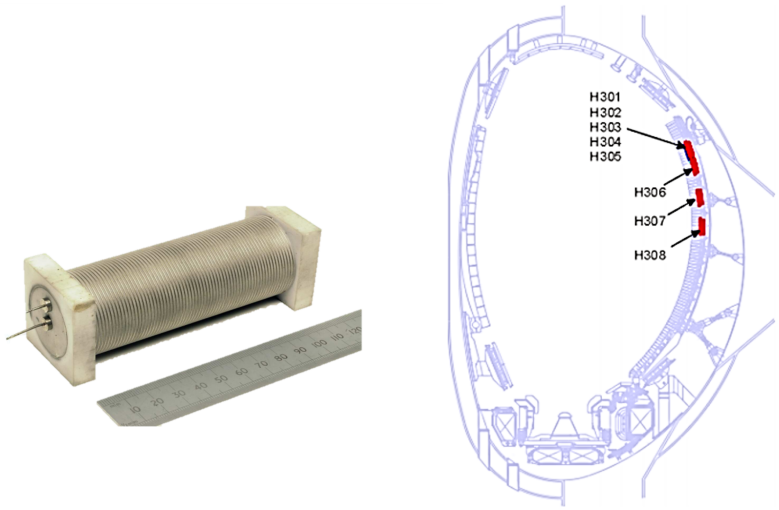


Figure 4.2: Example of a high-resolution array coil (left) and location inside the vacuum vessel at the poloidal cross-section view (right). Source: [81]

Figure 4.3 shows the magnetic pick-up coil data of the H305 coil from the database discharge 92213. The disruption instant is visible with a well defined spike, indicated by the dashed line. It is clear that the signal oscillates considerably throughout time, with alternating peaks. Due to this behavior, the use of processing techniques for non-stationary signals are more suitable to capture the frequency component.

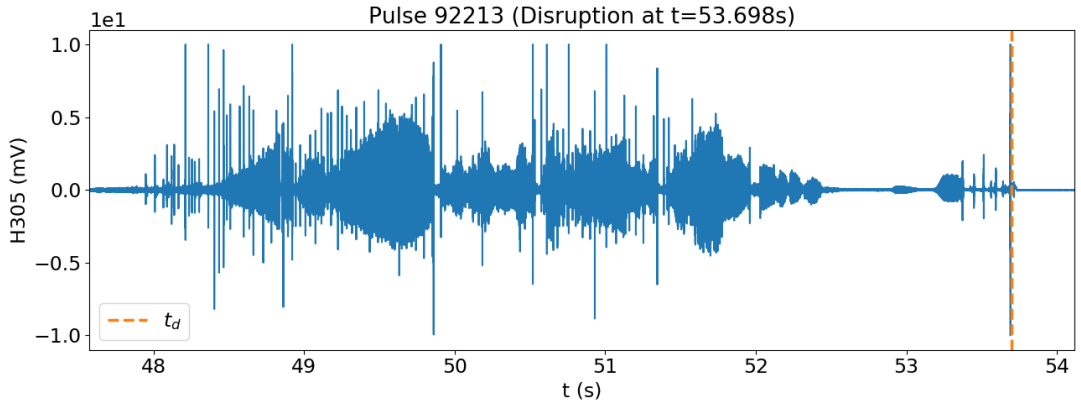


Figure 4.3: Raw data from the H305 coil. The dashed orange line indicates the disruption instant.

4.2.2 Locked Mode Amplitude

Of equal importance to this work is the measurement and use of the locked mode amplitude. The analysis of this data is included in the first part of the adopted methodology. Its measurement is possible due to a set of saddle flux loops (figure 4.4), which are located in the same poloidal plane at different radial positions, outside the vacuum vessel. These diagnostics are especially suitable for detecting non-rotating instabilities, such as locked modes, as at higher frequencies the first wall of the device could significantly shield the normal component of the magnetic field.

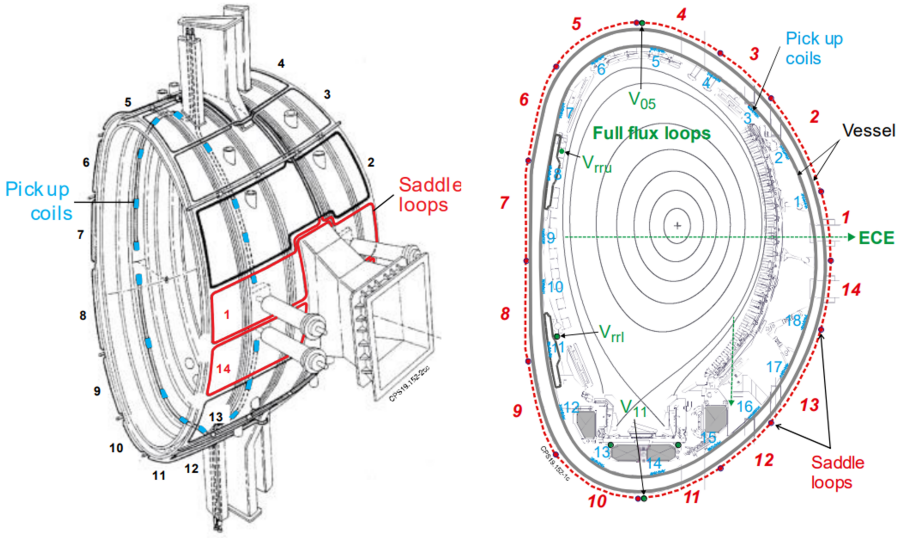


Figure 4.4: External vessel (left) and poloidal cross-section (right) views of the saddle flux loops. Source: [46]

According to reference [70], the amplitude of a $n = 1$ locked mode, B_{LM}, can be measured from the combination of magnetic field fluxes generated by various current sources of the setup and expressed as the contribution of both the sine B_{rSIN} and cosine B_{rCOS} components of the mode, as in equation 4.3.

$$B_{LM} = \sqrt{(B_{rSIN})^2 + (B_{rCOS})^2} \tag{4.3}$$

Prior approaches sometimes define a normalized locked mode amplitude to the plasma current [46, 50]. However, in the context of the addressed approach, normalization is discarded.

Examples of the locked mode signal from both disruptive and non-disruptive discharges can be seen in figure 4.5. As with the H305 data, a spike can be observed at the disruption instant, at $t = 53.70$ s. An increase in the locked mode amplitude at around $t = 53.40$ s in discharge 92213 is also noticed. This increase is relevant in physical or operational threshold-based methods to trigger an alarm. The initial exploration of the used database tells us that this did not happen with sufficient warning time in all disruptive discharges, in some cases even below the 30 ms benchmark. An example of the mentioned observation can be seen in figure 4.6, where the threshold is practically exceeded in the disruption

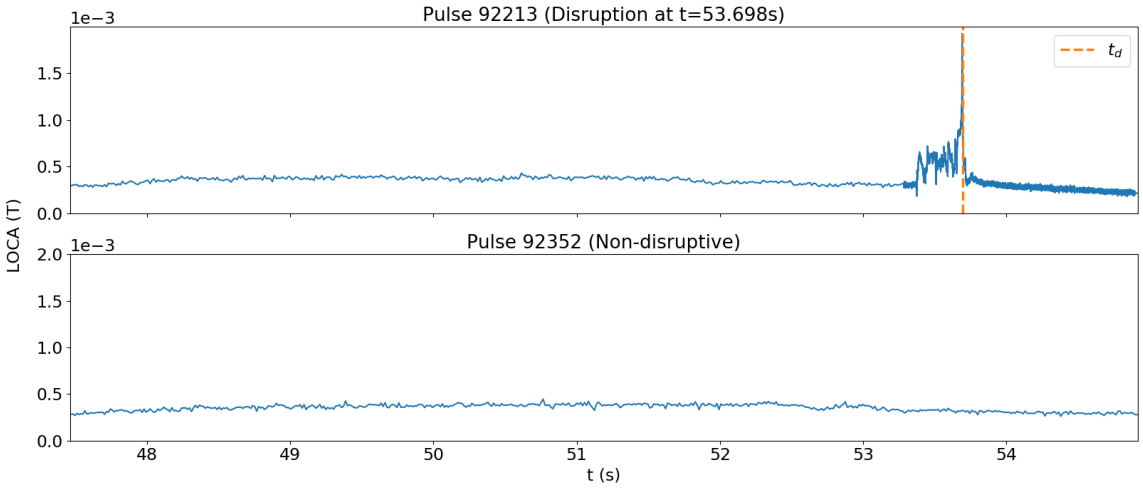


Figure 4.5: Example of the locked mode signal in a disruptive (top) and non-disruptive (bottom) discharge.

instant. This is a demonstration of possible limitations with the threshold approach when using it for the activation of the DMS, taking into account the ITER requirements.

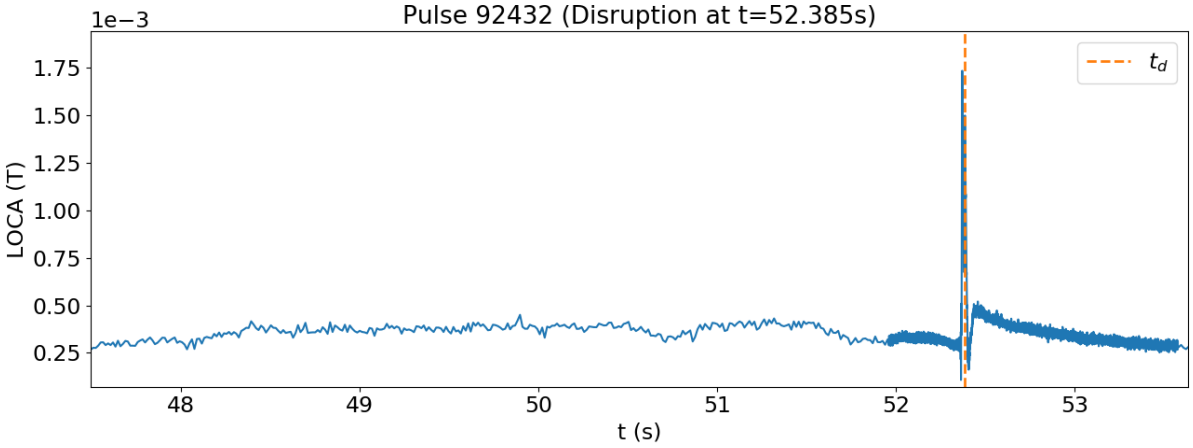


Figure 4.6: Disruptive discharge with a maximum in B_{LM} close to the disruption.

4.3 MHD spectrograms

The key instrument to analyze the MHD activity in this work is the spectrogram, a time and frequency representation of a given signal using Fourier analysis, in this case with data coming from the H305 Mirnov coil. More specifically, the used spectrogram is calculated using a discrete-time implementation of the short-time Fourier transform (STFT), a proper technique to be employed in non-stationary signals.

The process of the STFT can be computationally applied as follows: an input signal is divided into segments with a certain length N (which can be overlapped by a distance H from the segment center

to avoid artifacts), by multiplying it with a given moving window function $w(m)$. A Fast Fourier transform (FFT) is applied within each segment. The result is expressed in equation 4.4,

$$X(n,\nu) = \sum_{m=0}^{N-1} x(m+n)w(m)e^{-i\nu m} \tag{4.4}$$

The chosen window function $w(m)$ is a particular case from the generalised cosine window (Hamming function), with α = $25/46$, and defined in equation 4.5. A more detailed explanation of the STFT implementation can be seen in [82].

$$w(m) = 0.5\left(1 - \cos\left(\frac{2\pi}{N}m\right)\right) \tag{4.5}$$

To obtain the spectrogram representation, one must compute the magnitude squared of equation 4.4, $\left|X(n,\nu)\right|^2$. The spectrogram can thus be expressed as a 2D matrix, $S_{[t,\nu]}$, taking into account the global contribution of each segment and providing a local identification of the FFT amplitude $s_{t,\nu}$ at each t and ν indexes, as seen in 4.6,

$$S_{[t,\nu]} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,j} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ s_{i,1} & s_{i,2} & \cdots & s_{i,j} \end{pmatrix} \tag{4.6}$$

where $i, j \in (t, f)$. Each experimental pulse of our database is provided with a given matrix S and thus a spectrogram representation.

The choice of the STFT parameters had into account an adequate and balanced resolution in both time and frequency domains. Their relation is expressed through equation 4.7. Increasing the number of points in the segment rises the frequency resolution and removes temporal information. The opposite follows if one decreases the segment length.

$$\Delta\nu\Delta t \geq 1 \tag{4.7}$$

To compute the STFT, a segment length of 512 points was chosen. This was the same value applied for the number of overlap points. Using equation 4.7, and for a given sample frequency, $\Delta\nu = f_s/n$, the time resolution for this segment length is equal to 4.096 ms. The computation of the STFT was implemented with a GPU-supported framework for time optimization. Also, the chosen number for the segment length could not be increased due to the computational limitations of the available hardware.

Discharge 92213 is again used as an example. Its spectrogram representation is displayed in figure 4.7. The obtained values were converted to a logarithmic scale to better represent the data with a colormap. The disruption instants can be well characterized as a sequence of one or more bursts, well defined in time and covering almost the full spectrum of frequencies. Regarding the observation of MHD activity, the dynamics of mode-locking can be visible within this discharge. At $t = 51.5$ s, a $n = 1$

internal kink mode over a $q = 1$ surface starts to decrease in frequency, corresponding to the plasma deceleration. After an interruption in the MHD activity, a $n = 1$, $m = 2$ mode ($q = 2$ surface) appears at a lower frequency after $t = 53.0$ s. Its frequency is reduced and the mode eventually locks at around $t = 53.4$ s.

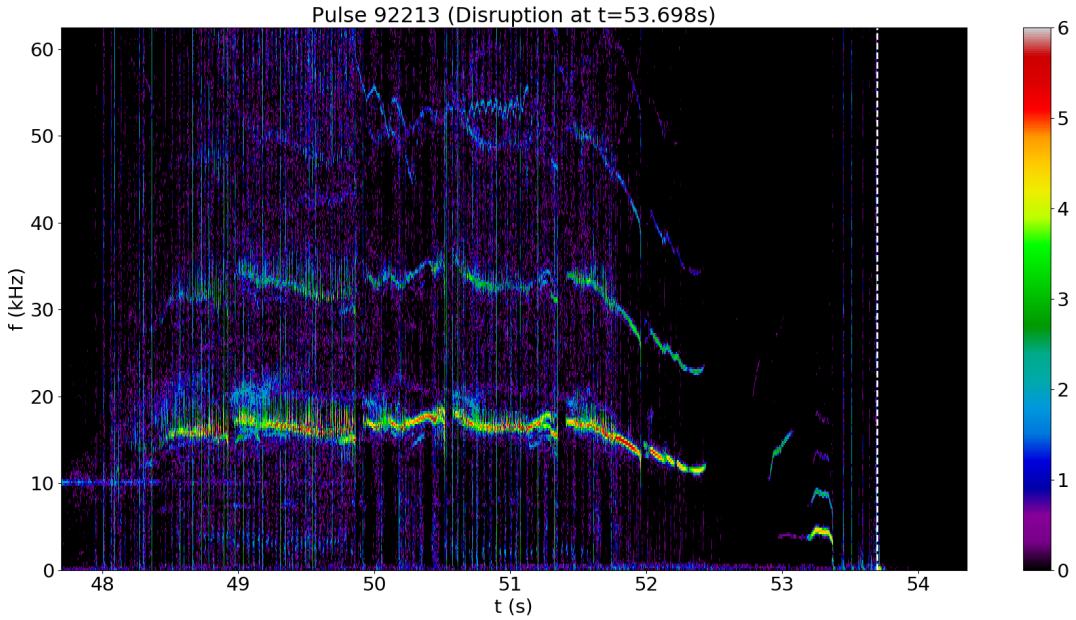


Figure 4.7: Spectrogram of discharge 92213. The dashed white line indicates the disruption instant.

4.4 Conclusion

In this section, we provided the necessary instruments to proceed with the proposed approach of this work. The spectrogram, for instance, is the main input source to our deep learning model, while the locked mode amplitude will serve as the main definition of disruption, substituting the previous label given by the database.

As a side note, although there are other methodologies for the study of non-stationary plasma signals [83], which can be more effective in dealing with possible resolution artifacts and with the frequency-time product (equation 4.7), only the STFT is considered in this framework to build the spectrogram.

5

Proposed Approach

Contents

With the experimental setup properly introduced, it is necessary to explain how the mentioned signals are used for this study, which is the goal of the present chapter. First, a brief description of the chosen data of JET is presented. A reformulation on the definition of disruption instant is also given, namely how the locked mode signal is used on this task. A deep learning model is developed in order to receive a spectrogram sample of a given pulse and to predict either if it is going to disrupt or not.

Finally, we use an interpretability tool to localize discriminative regions in the spectrogram that the model retrieves to make a certain prediction, giving some hints of possible disruptive MHD activity.

5.1 Dataset

The used data is comprised of a total of 486 pulses, ranging from JET campaigns C35 to C40, carried from 2015 to 2021. During this period, the experiments were framed in order to achieve a reliable and efficient steady-state D-T plasma at considerable confinement times τ_E (baseline scenario) [74], with a special focus on ITER conditions. In short, higher current and magnetic field profiles are achieved, with

a normalized pressure $\beta_N \sim 1.8$ and a safety factor at 95 % of the flux $q_{95} \sim 3$. Additionally, it holds a higher fraction of thermonuclear neutrons of the total neutron production when compared with the hybrid scenario due to less effective neutron beam injection. Figure 5.1 shows the performance of the baseline scenario compared with other experimental setups during 2016 campaigns [74].

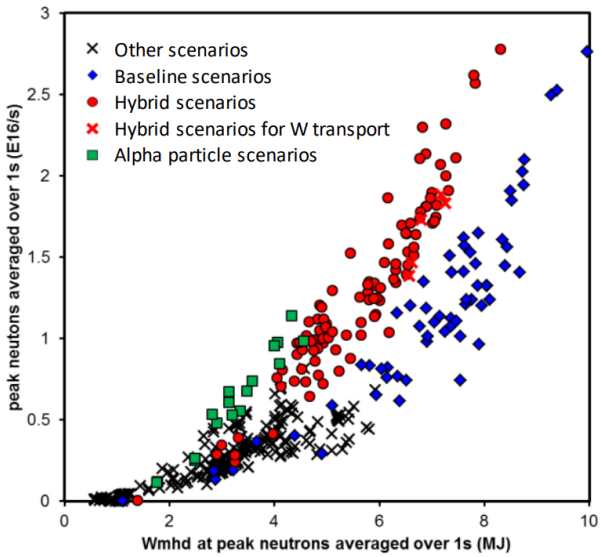


Figure 5.1: JET performance for baseline, hybrid and alpha particle experimental scenarios. Source: [74]

According to the prior database definition of disruption time, there are 291 non-disruptive and 195 disruptive pulses, thus a considerably balanced dataset in terms of the study target. However, as the first step, the study of a direct link between disruptive pulses and the occurrence of mode-locking was made, similarly to other studies, thus leading towards a new definition of a disruptive experiment in the context of this work.

5.2 Use of the locked mode amplitude

As already mentioned, the locked mode amplitude is widely used in terms of the machine operation, either for implemented disruption prediction systems, such as the APODIS, or in DMS valves. In both cases, the working principle depends on the monitoring of the signal and on the possible exceeding of a certain threshold, which may result in the activation of the alarm system. In the context of this work, the main focus is to understand how well can we split the available database between disruptive and non-disruptive experiments, by finding an optimal threshold value.

The initial study was framed as a binary classification problem using machine learning standards, however, in this stage, instead of having a trained classifier to make the predictions for our data, we attribute in advance the classification labels to each experiment. That is, each experiment was labeled

according to locked mode amplitude and its maximum value by using the following definitions:

- if the pulse is disruptive and the locked mode amplitude exceeds the threshold before the disruption time defined in the database, then the pulse is classified as a true positive (TP) event;
- if the pulse is non-disruptive according to the database definition and the locked mode amplitude does not exceed the threshold at any point in time, then the pulse is classified as a true negative (TN) event;
- if the pulse is non-disruptive according to the database definition and the locked mode amplitude exceeds the threshold at any point in time, then the pulse is classified as a false positive (FP) event;
- 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 defined in the database, then the pulse is classified as a false negative (FN) event.

Deciding the optimal threshold is the same as obtaining the best combination of previously chosen performance metrics which are commonly used for binary classifiers to a certain range of possible thresholds. The considered metrics are displayed in table 5.1. Due to its superior effectiveness with not complete balanced datasets, the Matthews correlation coefficient was considered as a complementary metric [84].

Metric	**Definition**
Accuracy	$\frac{TP+TN}{TP+TN+FN+FP}$
Balanced accuracy	$\frac{\frac{TP}{TP+FN}+\frac{TN}{TN+FP}}{2}$
F1-Score	$\frac{2TP}{2TP+FP+FN}$
Matthews correlation coefficient	$\frac{TP\times TN-FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Table 5.1: Binary classification performance metrics.

An optimal threshold value of 1.069×10^{-3} T was obtained when these metrics were applied to the proposed dataset (see figure 5.2). This value is in line with previous studies on DMS systems that use the locked mode [48], with 2×10^{-3} T, and the fact that most of these metrics values are above 95% shows a quite satisfactory discrimination between non-disruptive and disruptive experiments. Despite these results, and as seen previously, setting only a threshold value is somewhat limited because the warning time is relatively short in some disruptive discharges (the mean warning time in true positive events with the locked mode threshold is 75.3 ms). The obtained performance metrics values for the optimal threshold can be seen in table 5.2.

These first results are used to elaborate the following step, in which a binary classifier based on a deep learning model is developed to receive a sample of an MHD spectrogram and to assign it with a 0

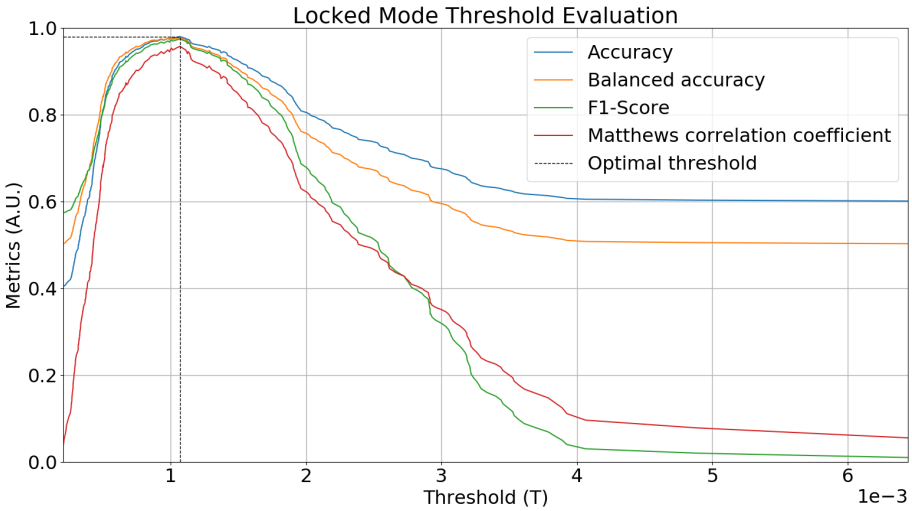


Figure 5.2: Binary classification metrics with a threshold range on the locked mode amplitude.

value if the locked mode amplitude does not exceed the 1.069×10^{-3} T threshold, and with a 1 value if otherwise.

Metric	Result
Accuracy	0.979
Balanced accuracy	0.977
F1-Score	0.974
Matthews correlation coefficient	0.957

Table 5.2: Performance results for the optimal locked mode amplitude threshold.

5.3 Training a neural network model

As stated in chapter 3, it has been of special interest to map a set of input variables from tokamak data to make disruption predictors. Some of the more recently developed predictors are built on top of deep learning models, namely with neural networks. Neural networks are based on two main objects: the neurons, which are mathematical objects that transform a given input, and blocks of layers, each one consisting of a given number of neurons.

Figure 5.3 shows an example of an artificial neural network (ANN). The first layer receives a given input vector $\mathbf{x} = \{x_1, x_2, ..., x_i\}$, while the neurons of a hidden layer K set the transformation of $\mathbf{x}$, y_i^K, using an activation function σ. This is expressed in equation 5.1, where w_i^K (weight) and b_i^K (bias) are

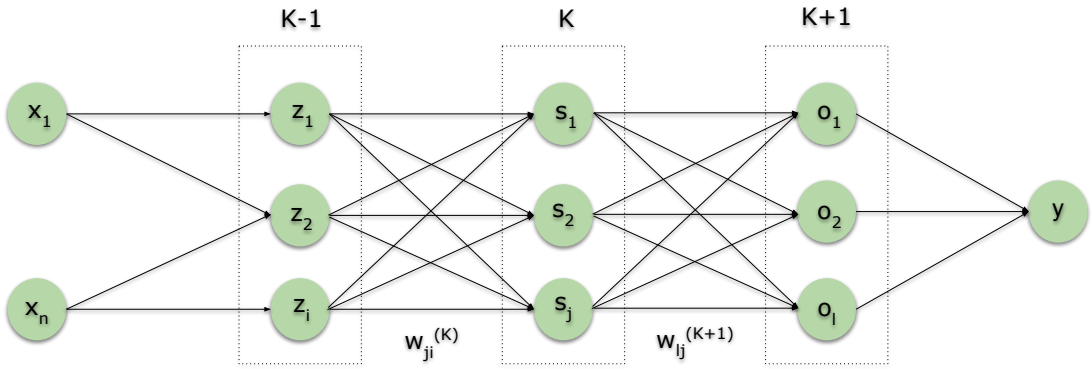


Figure 5.3: Schematic of a artificial neural network (ANN) with three hidden layers.

the parameters of the model in the hidden layer K.

$$y_i^K = \sigma(x_i \cdot w_i^K + b_i^K) \tag{5.1}$$

The calculated outputs of a given layer are directed towards the next layer. Finally, the output layer provides the prediction for a given value in the input layer.

Training a neural network is the same as optimizing the mapping between input and output values (targets) by tuning its parameters. To accomplish this, one must also define a loss function L that measures the difference between the previously labeled examples and the predicted values of the model. Each parameter from the network must be updated to minimize L and improve the ability of the model to generalize the targets. However, values close to absolute zero must be avoided - overfitting - since the probability of the model to predict correctly on new data becomes diminished. The definition of L depends on the type of classification problem one tries to solve.

It is also necessary a framework that controls how the parameters are updated to minimize L. This is done by an optimization algorithm that iteratively searches the best combination of parameters. One of the most known algorithms used for this purpose is the Gradient Descent. This algorithm searches for a convergence on a global minimum of L according to the gradient of the parameters of the model and the direction of the slope of its derivatives (see figure 5.4). Assuming that L is a differentiable function, each iteration of the gradient descent updates a parameter θ_{n+1} according to the updated rule given in equation 5.2,

$$\theta_{n+1} = \theta_n - \nabla L(\theta_n) \tag{5.2}$$

where θ_n is the parameter value given in the previous iteration and η is the learning rate. Both the size and speed of each iteration are determined by the hyperparameter η. Variations of the gradient descent were also developed to improve convergence.

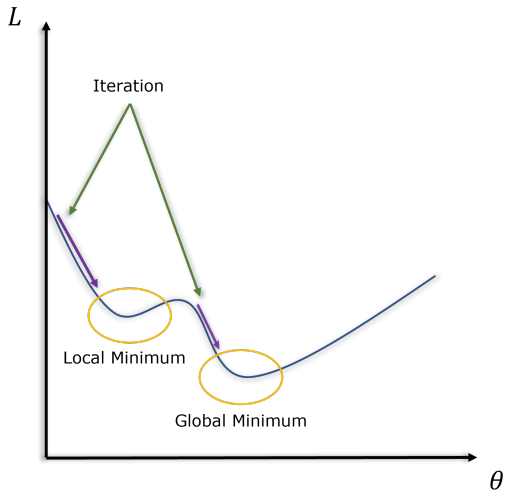


Figure 5.4: Illustration of the Gradient Descent algorithm.

To transport the gradients of L from Gradient Descent through the different layers of the network, from the last to the first layer, one can use the backpropagation technique. Using figure 5.3 to illustrate backpropagation, the network is first feed-forward with the predictions for a given input vector. That is, we compute the output of a given unit in a given layer, $s_j^{(k)}$, using the units from the previous layer, $z_i^{(k-1)}$, as in equation 5.3,

$$s_j^{(k)} = \sigma\left(w_{j0}^{(k)} + \sum_i w_{ji}^{(k)} z_i^{(k-1)}\right) \tag{5.3}$$

where $w_{j0}^{(k)}$ is the weight associated with the unit 0 of a layer k and $w_{ji}^{(k)}$ is the weight from the unit i in the same layer. After obtaining the prediction $\hat{y}$, the gradient of L needs to be calculated and backpropagated to the network. Thus, the gradient of L with respect to unit $s_j^{(k)}$, for example, is calculated according to equation 5.4,

$$\frac{\partial L}{\partial s_j^{(k)}} = \sum_l \sigma'(o_l^{k+1}) w_{lj}^{(k+1)} \frac{\partial L}{\partial o_l^{(k+1)}} \tag{5.4}$$

where $\sigma'(o_l^{k+1})$ is the derivative of the activation function of unit l with respect to its input. Similarly, we do the same for each weight that constitutes $s_j^{(k)}$, as in equation 5.5.

$$\frac{\partial L}{\partial w_{ji}^{(k)}} = \frac{\partial L}{\partial s_j^{(k)}} \sigma'(s_j^{(k)}) z_i^{(k-1)} \tag{5.5}$$

The parameters of the network are updated using the relation in 5.2 when the gradients of L are computed for all the units. In practice, the gradient of L is sequentially calculated on different groups of training examples (batch) until the entire set is passed through the network (when this happens it is called a training epoch). The training process of a deep learning model with supervised learning could thus be summarized as the fitting of the model to training data and continuously trying to minimize the

difference between real and predicted values by adjusting its parameters with a set of validation data.

Despite the ability of fully connected neural networks to solve non-linear tasks, one finds some disadvantages, also taking into account the objectives of the proposed work. The simplest network is the single-layer perceptron [85], with only input and output layers, mainly directed towards binary classification. The feature retrieval is relatively easy in these networks since to each input only a given weight is associated. One could detect, for instance, which part of the spectrogram contributed more to the classification. However, the network is also limited when the complexity of the input increases, thus it would not be effective in linearly separating the spectrogram samples to the non-disruptive and disruptive domains.

Additionally, two main constraints arise if one increases the number of hidden layers and the model size. Possible interpretability from multi-layer perceptrons becomes complicated because the relation with each input and the model parameters becomes non-linear. Also, to gradually increasing input sizes, training becomes more computationally challenging since the number of connections between nodes and layers also increases, as well as the number of parameters [86].

Convolutional neural networks (CNN) are a specific class of neural networks, particularly probed to deal with image classification tasks [87], and a valid alternative to fully connected layers when treating spectrogram data [88]. Mainly two additional types of layers are added with respect to the ANN. In convolutional layers, the convolution product between the input and a filter matrix (also called the kernel) is computed, i.e., a dot product between input and filter matrices, with a moving window that slides at different spatial locations of the input, to produce a lower representation from the original source. This representation is called a feature map, which, in an image, can reflect some relevant changes in the geometry or lines of sight learned by the network. Mathematically, both the convolutional product in a given convolutional layer k and the obtained feature map C_k, can be defined with the dot product between a 2D input I and a 2D moving kernel K with equation 5.6 [90],

$$C_k(u,v) = \sum_m \sum_n I(m,n) \cdot K(u-m, v-n) \tag{5.6}$$

The convolutional layers also depend on additional hyperparameters that can be adjusted. The most common ones are the size of each filter/kernel in width and height, the spacing between consecutive filters in the input (stride), and the padding, which is applied to preserve the tensor information at the ends of the input matrix.

After the generation of the feature maps in the convolutional layer, these can be passed to a pooling layer, where downsampling is performed. This downsampling can be useful to avoid the learned features to considerably fluctuate due to small changes in the input, as well as to reduce the amount of computation. From the different types of pooling layers, the max-pooling is one of the most used, in which the maximum of each component from the feature map C_k is retrieved, as seen in equation 5.7.

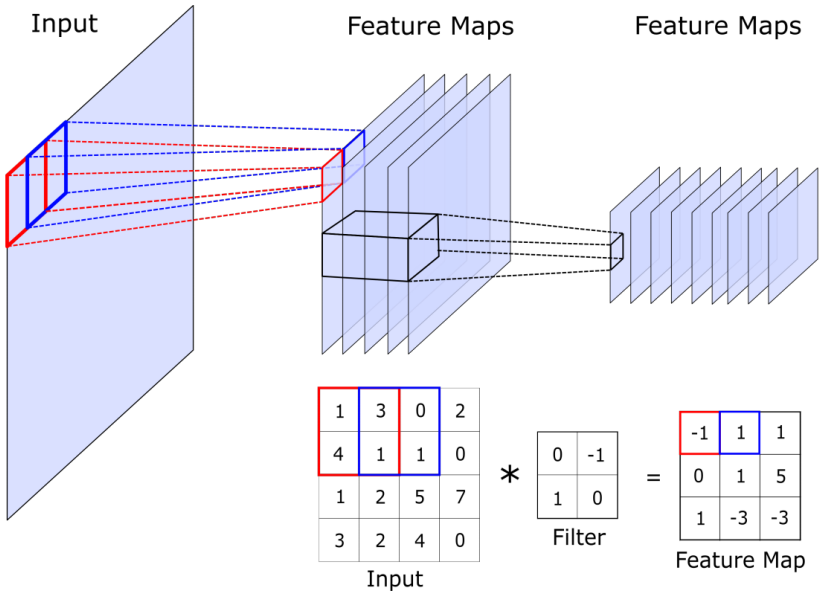


Figure 5.5: Convolutional product demonstration with a filter window and the feature map calculation. Source: [89]

Figure 5.6 reflects the working of a max-pooling layer. A more complete description of the CNN can be seen in [90].

$$P_k = \max(C_k) \tag{5.7}$$

Also, of particular importance for this work is the use of a global average pooling layer. In this layer, each feature map from the previous convolutional layer is averaged, flattening the input to a 1D vector. The resulting 1D vector can then be fed either to the fully connected layer or to an activation function to obtain the probability distribution of the classifier.

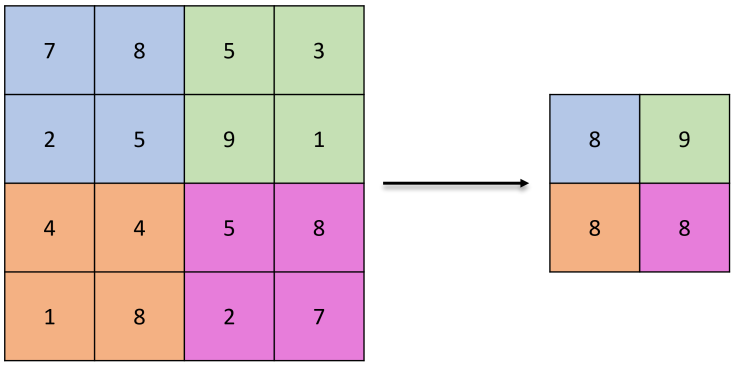


Figure 5.6: Representation of the max pooling, where a 2x2 stride on the feature map (left) selects the maximum values (right).

5.4 Proposed Model

The proposed model is based on the CNN described in the previous section and can be seen in figure 5.7. As indicated in the red region, at the beginning of the network, the input consists of sample windows from the spectrogram (previously computed from the Fourier analysis) of 256 to 256 points in both the time and frequency axis. This means that it covers approximately 1.049 seconds in the time axis and the whole frequency axis. In each 2D convolutional layer, the two initial values refer to the width and height of the resultant feature maps, while in each max-pooling layer it refers to each feature map size after downsampling. The third term is the number of applied kernels to the input in the convolutional product and, for instance, the number of obtained feature maps.

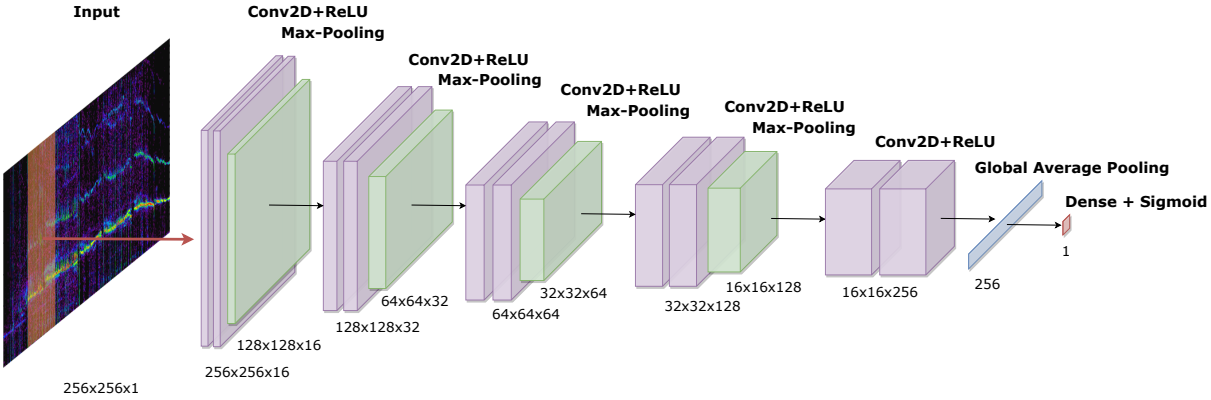


Figure 5.7: Proposed CNN model.

Following the input layer, the model is composed of five convolutional blocks, each block with two consecutive 2D convolutional layers. These consecutive layers allow a gradual increase in the complexity level of the features retained by the kernels, with the lower representations obtained in the first convolution, which are then fed to the second convolutional layer to obtain the most relevant components. A rectified linear unit (ReLU) activation function is used within the convolutional layers to positively normalize its inputs and to prevent the full stack of neurons to be activated [91]. It is defined in equation 5.8.

$$ReLU(x) = \begin{cases} 0 & if\ x < 0 \\ x & if\ x \geq 0. \end{cases} \tag{5.8}$$

A max-pooling layer is placed after the two convolutional layers, except at the last block. Due to the implementation of the interpretation technique, the max-pooling layer is substituted with a global average pooling layer, which makes the spatial average of the 256 feature maps coming from the last convolutional block to a 1D vector of 256 values. Experiments done during the implementation of our model showed that substituting this layer with two dense layers does not lead to a significant divergence of the training results.

To conclude, the resulting vector from the global average pooling layer is fed to a one unit dense layer with a sigmoid activation function (equation 5.9) to provide the probability value between 0 and 1.

$$f_{sigma}(x) = \frac{1}{1+e^{-x}} \tag{5.9}$$

The model is trained to discriminate two classes (non-disruptive or disruptive sample due to the locked mode), thus the binary cross-entropy L function will be applied in this work to measure the difference between the real training sample labels and the sigmoid function values of the output layer. It can be described as equation 5.10,

$$L(\hat{\boldsymbol{p}}, \boldsymbol{p}) = -\frac{1}{\mu}\sum_{j=1}^{\mu}[p_j \ln(\hat{p}_j) + (1-p_j)\ln(1-\hat{p}_j)] \tag{5.10}$$

where p_j is the real binary label of a training sample, $\hat{p}$ is the predicted output and μ the total number of training samples. Moreover, we used a variation of the Gradient Descent algorithm, called Adaptive Momentum Estimator (Adam) [92]. To apply the Adam optimizer, equation 5.2 is modified and given by equation 5.11, introducing new hyperparameters.

$$\theta_{n+1} = \theta_n - \frac{\eta}{\sqrt{\frac{v_n}{1-\beta_2^n}} + \epsilon}\frac{m_n}{1-\beta_1^n} \tag{5.11}$$

In equation 5.11, ϵ avoids the fraction to be divided by 0, while β_1 and β_2 are decay rates of first and second order, respectively. m_n and v_n are called the first and second momentum of the function's gradients, and defined as equations 5.12 and 5.13.

$$m_n = \beta_1 m_{n-1} + (1-\beta_1)\left(\frac{\partial L(\theta_n)}{\partial \theta_n}\right) \tag{5.12}$$

$$v_n = \beta_2 v_{n-1} + (1-\beta_2)\left(\frac{\partial L(\theta_n)}{\partial \theta_n}\right)^2 \tag{5.13}$$

The model has been trained using samples of 485 discharges, where one of them (92213) has been left out for testing purposes. The input window could be sampled anywhere on the time axis of the spectrogram, i.e., at the beginning of the experiment, where there is not any visible evidence of the possible outcome, and sometimes a very low frequency regime with no MHD activity, or after the disruption instant in a disruptive pulse, if the locked mode has already occurred. We also allowed the window to be sampled anywhere in order to avoid any possible bias in the model towards a desired outcome.

90% of the total discharges were used as training set, and 10% as a validation set. Each training batch contained exactly one sample from each training experiment so that the batch size was equal to the number of training pulses. Additionally, at each batch, a different sample is drawn from the spectrogram

of each training pulse to avoid any repetition. At the end of each epoch, we took 100 equally distributed samples from each of the 10% set experiments for validation. As a key note, all of the time axis of each discharge of our dataset is considered in both training and validation sets, i.e. more than one 256x256 sample window from each batch is considered.

It is also important to notice that no prior assumption was made regarding the time from which the samples are considered, namely because we do not have prior knowledge on any possible correlation between the frequency information of the H305 signal and other diagnostics, such as the plasma current or the NBI system.

5.5 Class Activation Mapping

In order to obtain physical interpretation from the classifier and possible relevant MHD activity, we used the class activation mapping (CAM) [68] method. With CAM, we get the discriminative localization that highlights the regions that contributed the most to the classification of a determined input by the CNN model. In an image, for example, the result can be interpreted as a heatmap, where the value of each pixel indicates how important that pixel is for a given classification. In the case of our input type, it determines the frequency region that the trained CNN model finds to be the most relevant to predict mode-locking disruptions.

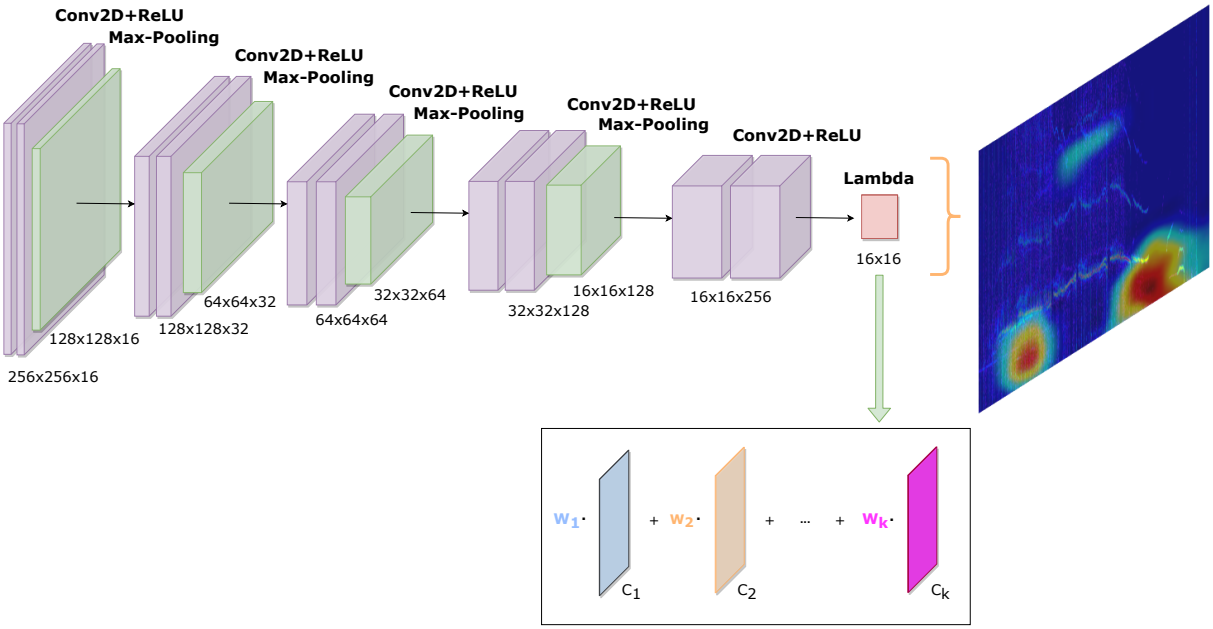


Figure 5.8: Model after the introduction of CAM. Note the substitution of the global average pooling layer from the previous CNN model.

As previously described, the global average pooling layer reduces each of the feature maps from the

last convolutional layer to 256 average values. These averaged values are then passed to a dense layer to produce the model prediction. However, in CAM, we compute a weighted sum of the features maps, each one associated with an equivalent dense layer weight, as seen in figure 5.8.

Since the feature maps are taken before any dense layer, this technique employs a slight change to in the perspective of the model. This change has been minimized in advance when implementing our model, as we were aware of the CAM constraints. As seen in figure 5.8, after training the model, both the global average pooling layer and the dense layer are removed and replaced with a single custom layer that computes the weighted sum of 16×16 feature maps produced by the last convolutional layer. In order to overlay the map on the input sample and be able to directly observe the results, it is also upsampled to the same dimensions of the input (256×256).

It may be argued that there is a possible trade-off between interpretability and the capabilities of the model to classify correctly when CAM is applied [93], which leads to the modifications of the original models. This is because of the complexity of the fully connected layers with extra parameters that can be added to CNN in order to improve the accuracy [94]. However, and as stated previously, the experiments made within the framework of this work have shown us that there are no relevant differences in terms of our model's accuracy with the addition of fully connected layers. For this reason, the trade-off can be ignored with a certain degree of confidence.

6

Results and discussion

Contents

Following the presented methodology, this chapter gives the main results obtained. First, the training and performance metrics of the implemented model are shown, as well as the used hardware for computational allocation and the hyperparameter selection for the optimal experiment. Furthermore, the application of CAM is also approached, where we seek physical insight.

6.1 Model Training

Both the CNN model and the CAM methods were implemented through Python code, using the TensorFlow backend library for architecture design and training with GPU computation support. The generation of the spectrogram sample data and the synchronous training of our model were done using two Nvidia TITAN X GPUs with CUDA compatibility. The hyperparameter selection for the optimal model training, after many trials, can be seen in table 6.1, and the results of training are shown in figure 6.1. At epoch 357, the best value of validation loss L_v is achieved (0.38), reaching a validation accuracy of 0.83. After achieving the global minimum, it is clear the development of overfitting, mainly with a considerable divergence between both the training loss L_t and L_v, this last one possibly reaching absolute zero if more

epochs were given. Another interesting point is the presence of alternate noise on these metrics, which is keener on L_v.

Hyperparameter	Value
Epochs	737
Batch size	437
Learning rate (η)	1×10^{-4}
First-order decay rate (β_1)	0.900
Second-order decay rate (β_2)	0.999
Numerical constant (ϵ)	1×10^{-7}

Table 6.1: Hyperparameter selection for the optimal training of the model.

These results could be attributed to the nature of both the input data and the method applied in training. As previously mentioned, no constraints were set when choosing a given sample in training, i.e., they could be selected at any time of the discharge. This means that, at the beginning of the discharge, the chosen sample is identical between non-disruptive and disruptive cases. Thus, if the behavior in frequency is practically the same in two samples, even with different labels, the model can behave very well on training data, but wrongly guess in test data.

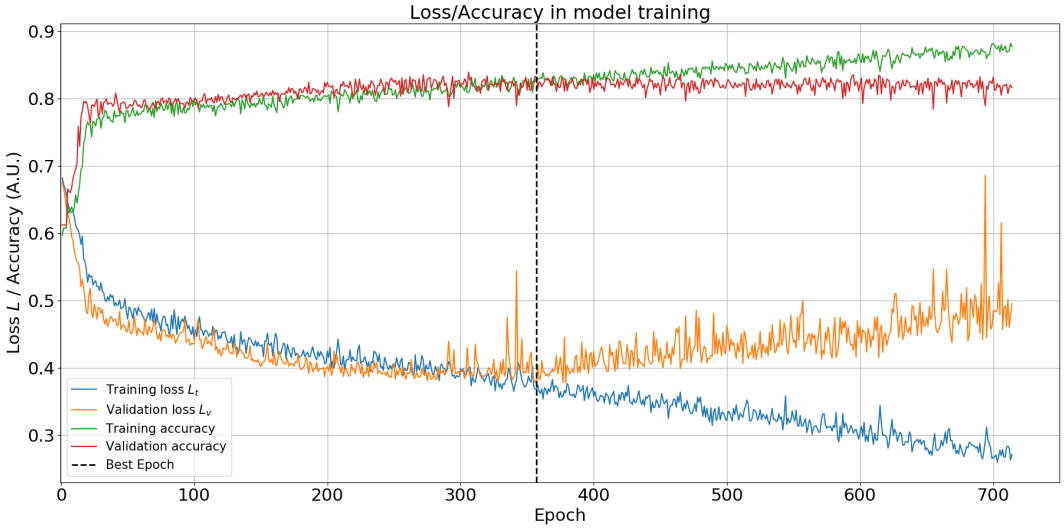


Figure 6.1: L function and accuracy metrics obtained during the optimal training.

This issue is commonly treated with the addition of regularization techniques [95]. However, even the addition of these techniques could not solve overfitting without sacrificing accuracy, which is important if one tries to understand the difference in mode-locking mechanisms with other samples.

6.2 Model Predictions

As stated, during training, a discharge was left out. With the model properly fitted with training data, we used discharge 92213 to see the prediction results. Each spectrogram sample window, at a given time index, was used to make a prediction for the next window, from the beginning of the discharge to the last sample. That is, the sample window is positioned always in the past in the time from which the prediction is made. The results are shown in figure 6.2.

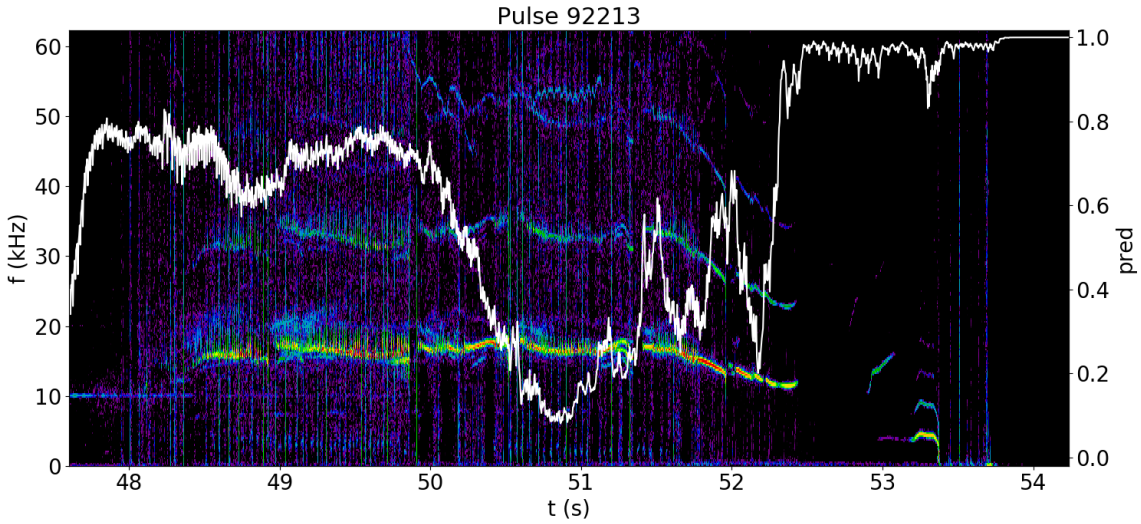


Figure 6.2: Model prediction (white) for the spectrogram samples of discharge 92213.

One of the first visible characteristics of this result is the considerable noise present along with the prediction values. Similarly to the fluctuations present with the training metrics, this can be due to the uncertainty of the model to classify these samples, as during training, it would be learning very resemblant samples with different labels.

A more relevant characteristic could be considered during the decrease in frequency of the $n = 1$ internal kink mode and before its interruption. At $t = 52.2$ s, the prediction continuously increases to almost 1, and practically stabilizes during MHD activity interruption (from $t = 52.5$ s to $t = 53$ s), mode-locking (from $t = 53.2$ s to $t = 53.4$ s) and the disruption instant. In practice, the model was able to recognize a disruption due to the locked mode even before its development in the spectrogram, with a time window of approximately 1 s. This is an interesting fact when the approach is compared with the locked mode amplitude, as in figure 6.3, where only at $t = 53.4$ s there is an increase in its value. Comparing this method with the APODIS system, where the average warning time is about 350 ms, the present approach, at least within this experiment, could provide an improvement in the time which the DMS would have to avoid the disruption.

An example of a non-disruptive discharge due to the locked mode can be seen in figure 6.4. In this

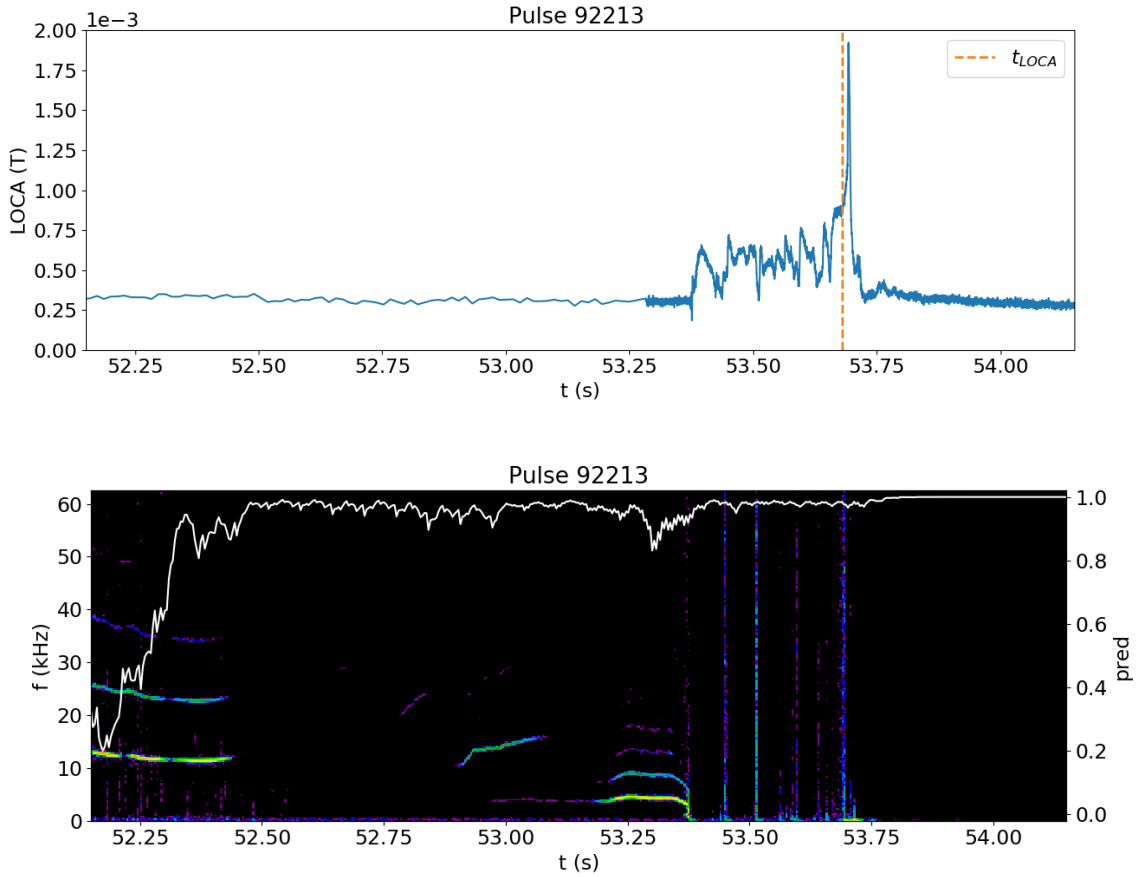


Figure 6.3: Comparison between the locked mode amplitude (top) and the developed predictor (bottom).

discharge, it is possible to verify the same fluctuations of the probability value given by the predictor as observed in discharge 92213. It becomes clear the similarity between the various samples of the two discharges, namely with a presence of internal kink modes. The model may not be able to discriminate a given sample between the two classes until the later phase of the discharge. The main difference between the two mentioned examples is the absence of mode-locking in discharge 92352, which makes the probability not to saturate at higher values.

6.3 Performance Metrics

One of the relevant tasks when developing binary classifiers is to test them in a new set of data and evaluate its performance metrics to measure its capability to generalize. This is also important when defining an optimal probability threshold that maximizes the performance metrics to be used as an alarm value when the predictor is implemented in a tokamak. However, almost all of the available data was used during the training of the model, thus a test set with new experiments could not be considered in

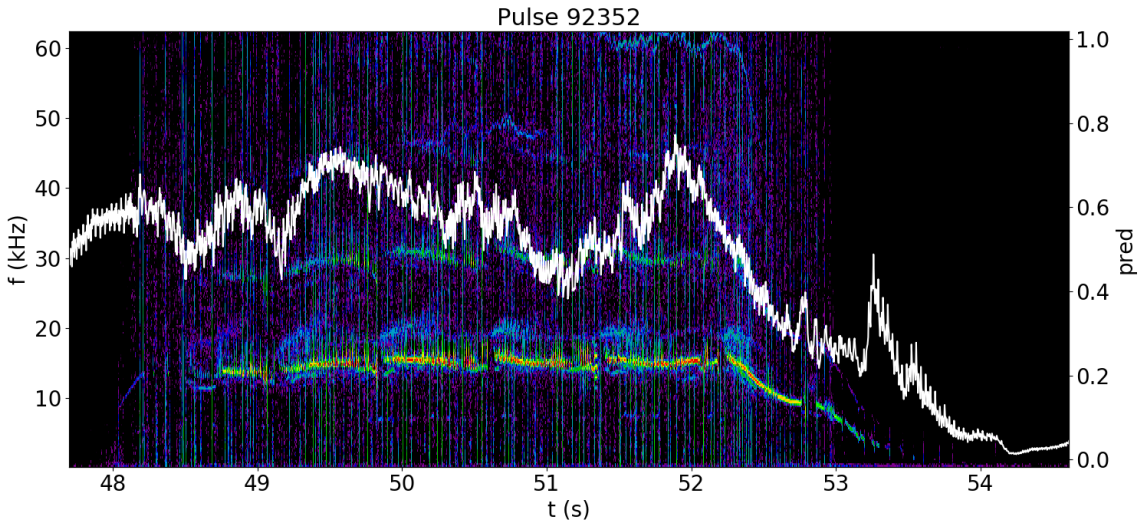


Figure 6.4: Model prediction (white) for the spectrogram samples of 92352 discharge (non-disruptive).

this work. Choosing a new set of JET discharges outside the baseline framework could limit any further conclusions with the model used since the experimental conditions are different.

Since the validation set has less influence on the model performance than the training set, it was used as an equivalent test set to elaborate this task, where 48 discharges were included. Similarly to what was described in chapter 5, if the probability does not reach a defined probability threshold value when the discharge has a locked mode disruption (surpasses the 1.069 mT mark in locked mode amplitude), then the discharge is classified as a FN event, and if it reaches that threshold in a 0 labeled sample is classified as a FP event. Also, the same performance metrics used in chapter 5 were considered, with the addition of the receiver operating characteristic (ROC) curve [96]. The results are shown in figure 6.5.

At the probability threshold value of 0.883 all the metrics reach their maximum, with an accuracy of about 79.5%. The reason for the considerably high value obtained for the probability threshold can be attributed to the fluctuations present during the predictions at both non-disruptive and disruptive discharges. At discharge 92213, it would be activated when there is an interruption of the MHD activity, at about $t = 52.5\ \mathrm{s}$. From all the metrics considered, the Matthews correlation coefficient seems to give the worst outcome of the model, as the result is practically between a random and a good classifier, according to the metric standards. The ROC curve indicates that the predictor is reasonably far from a random classifier, which is a good indication, confirmed by the area under curve (AUC) obtained (0.82), despite the fact that the number of FP discharges is still considerable. The associated confusion matrix of the classifier when applied to the validation set can be seen in figure 6.7.

Considering the optimal probability threshold given by the performance metrics, and when applied to the whole dataset, it is also relevant to point out the variation of classification compared with the

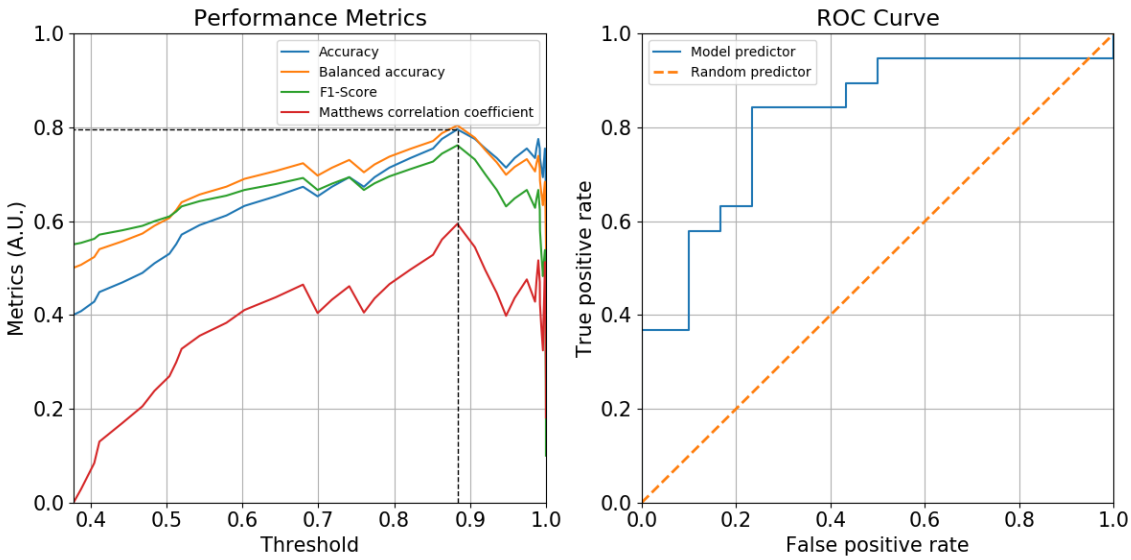


Figure 6.5: Performance metrics for different probability thresholds (left) and ROC curve (right) of the developed predictor.

locked mode amplitude threshold (see figure 6.6). More discharges are classified as disruptive with the developed predictor, which can be due to the large fluctuation of the probability throughout the experiment, falsely identifying non-disruptive discharges as disruptive in some cases.

Metric	Result
Accuracy	0.796
Balanced accuracy	0.804
F1-Score	0.762
Matthews correlation coefficient	0.594
AUC	0.819

Table 6.2: Best performance metrics achieved for a probability threshold of 0.883.

Having into consideration the limitations of the training data, and due to the fact that we are considering only one feature, the results obtained are still quite interesting. However, they cannot be directly compared with some of the past approaches for disruption prediction (see Chapter 3), due to the lower accuracy value, and the fact that no test set was used to measure the performance metrics. In addition, we are just taking into consideration MHD activity and the possibly related mode-locking mechanism, with other phenomena being left out of scope. Due to these facts, the comparison is beyond the scope of this work.

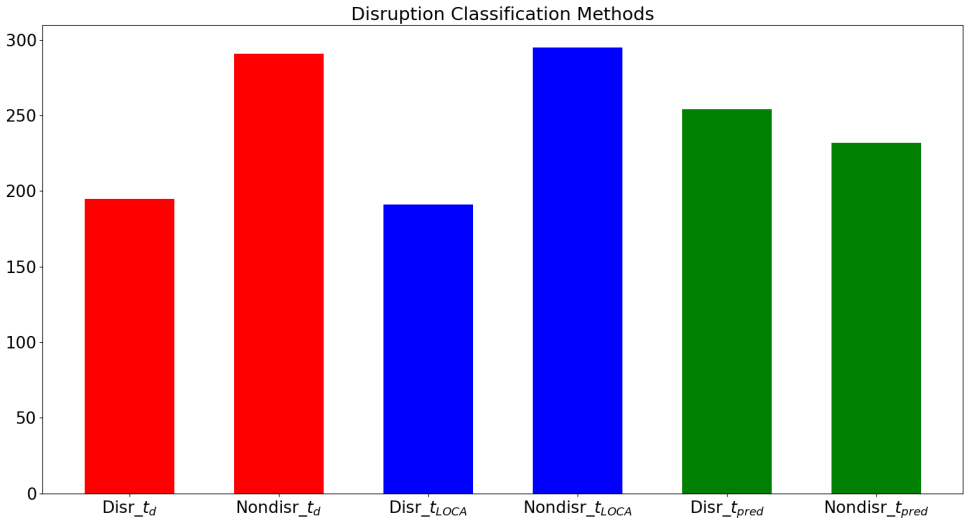


Figure 6.6: Classification comparison between the database disruption definition t_d (red), locked mode threshold t_{LOCA} (blue) method, and developed predictor t_{pred} (green).

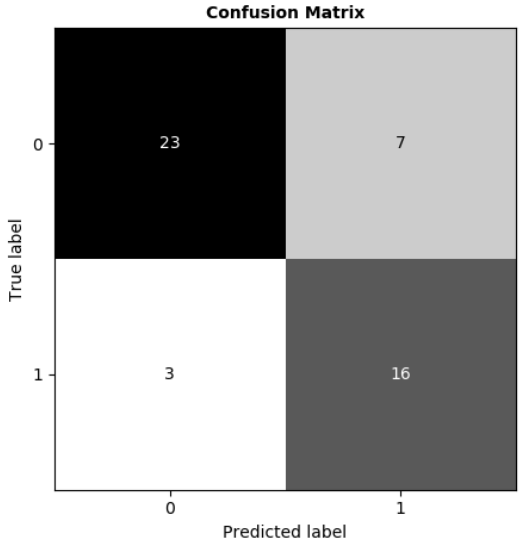


Figure 6.7: Confusion matrix from the application of the predictor in the validation set with a probability threshold of 0.883.

6.4 Interpretability with CAM

To obtain the features retrieved by the model, we used the modified model (see figure 5.8) which computes heatmaps for a given sample. The negative values (as blue color) represent the areas of the spectrogram that contributed the most to the classification of the predictor towards 0, which in the case

of a more intense blue indicates the saturation of the sigmoid function to its minimum value. For positive values (as red color), the areas in the spectrogram had a higher influence towards a classification of 1. The intermediate value (as white color) is equivalent to a random guess from the predictor. The resultant map is overlaid on the spectrogram sample to facilitate further interpretations. The application to discharge 92213 can be seen in figure 6.8.

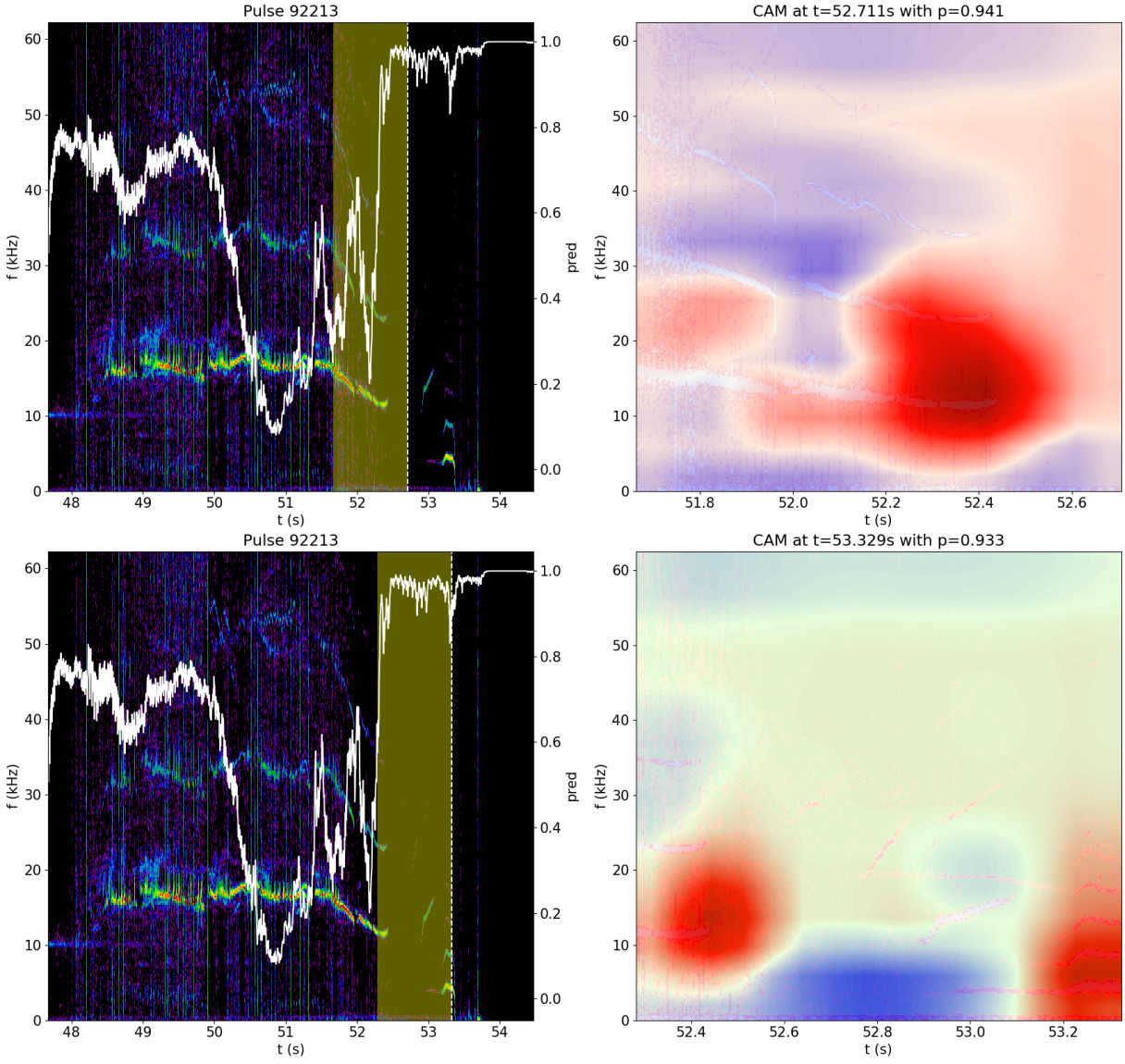


Figure 6.8: Result of CAM for discharge 92213 at $t = 52.71$ s (top) and $t = 53.32$ s (bottom).

The right-hand side of the image corresponds to the CAM method when applied to the sample highlighted by the yellow rectangle on the spectrogram at the left side. It can be observed what most influenced the predictor to make a certain classification. In the case of discharge 92213, there are features that converge with the physical interpretation given by the previous analysis of the spectrogram.

First, it is interesting to notice the highlight on the interruption of the internal kink mode, specifically at $t = 52.4\,\mathrm{s}$. This highlight by CAM is sufficient to considerably raise the prediction value to approximately 1, about one second before mode-locking development. Secondly, two additional characteristics can be seen by moving the sample window. A negative area, where practically no MHD activity is present (only a residual frequency band), and the positive highlight in a $m = 2$, $n = 1$ mode ($q = 2$ surface), at a lower frequency (approximately 4 kHz), which decreases due to locking afterwords. This kind of behavior is consistent with the work by Sweeney et al. [44], where these modes, namely rotating $m/n = 2/1$ modes, are likely to lock.

One could also argue that the window length of a given sample that is provided to the model has a certain influence on the overall results. In fact, the result in the bottom CAM of figure 6.8 could be slightly different if the latter $n = 1$ mode was not framed by the window, due to the weighted average of CAM. It is also interesting to observe the difference between the CAM highlights in the interruption (t_{CAMi}) and resurgence (t_{CAMr}) of the MHD activity, and the probability saturation of the predictor (t_{pred}) with the time in which the B_{LM} threshold is reached (t_{LOCA}). This can be seen in figure 6.9.

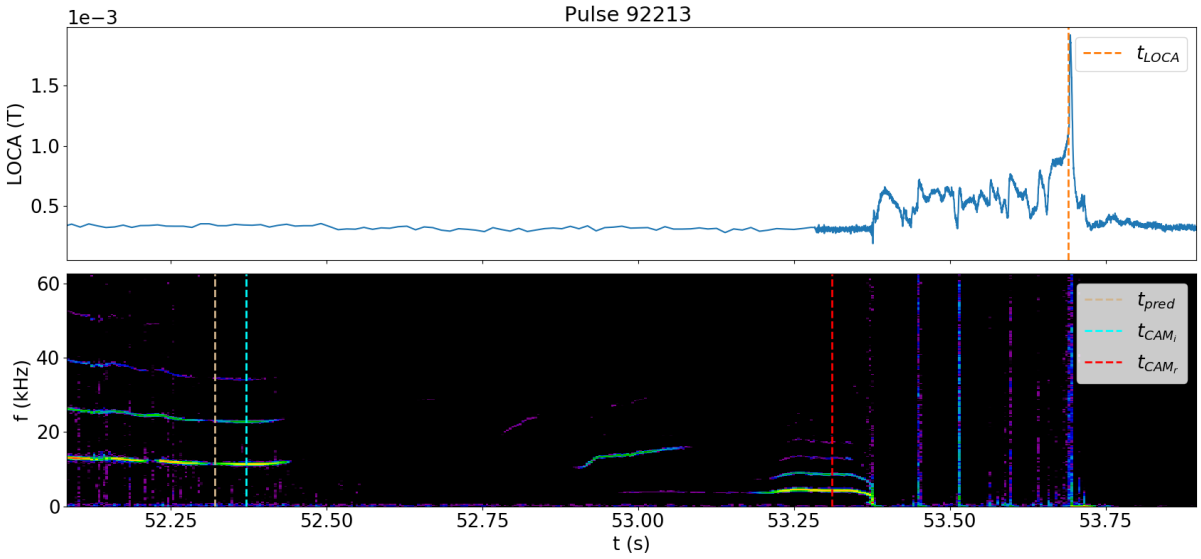


Figure 6.9: Locked mode threshold (top), CAM highlights, and probability threshold (bottom) at discharge 92213. The last two observations are before any major increase in the locked mode.

The same pattern can be observed when the method is applied to discharge 96996 (see figure 6.10), for example. As in the previous case, the CAM method gives particular interest to the suppression of both $n = 1$ and $n = 2$ modes, at approximately 10 kHz and 20 kHz, respectively. This is also where the probability starts to increase, providing a warning of a possible disruption due to the locked mode. Moving the yellow window in time, as a $m = 2$, $n = 1$ mode appears at lower frequency (approximately at 5 kHz), again the CAM method seems to give considerable relevance. It also includes the continuous

decrease in the frequency of the mode at $t = 57.08$ s. In Pucella et al. [97], this is also observed and correlated with the growth and destabilization of a $m/n = 2/1$ island due to temperature hollowing, in which some impurities start to accumulate in the plasma core. The suppressed mode at $t = 56.4$ s is also associated with a $n = 1$ mode at the $q = 1$ surface.

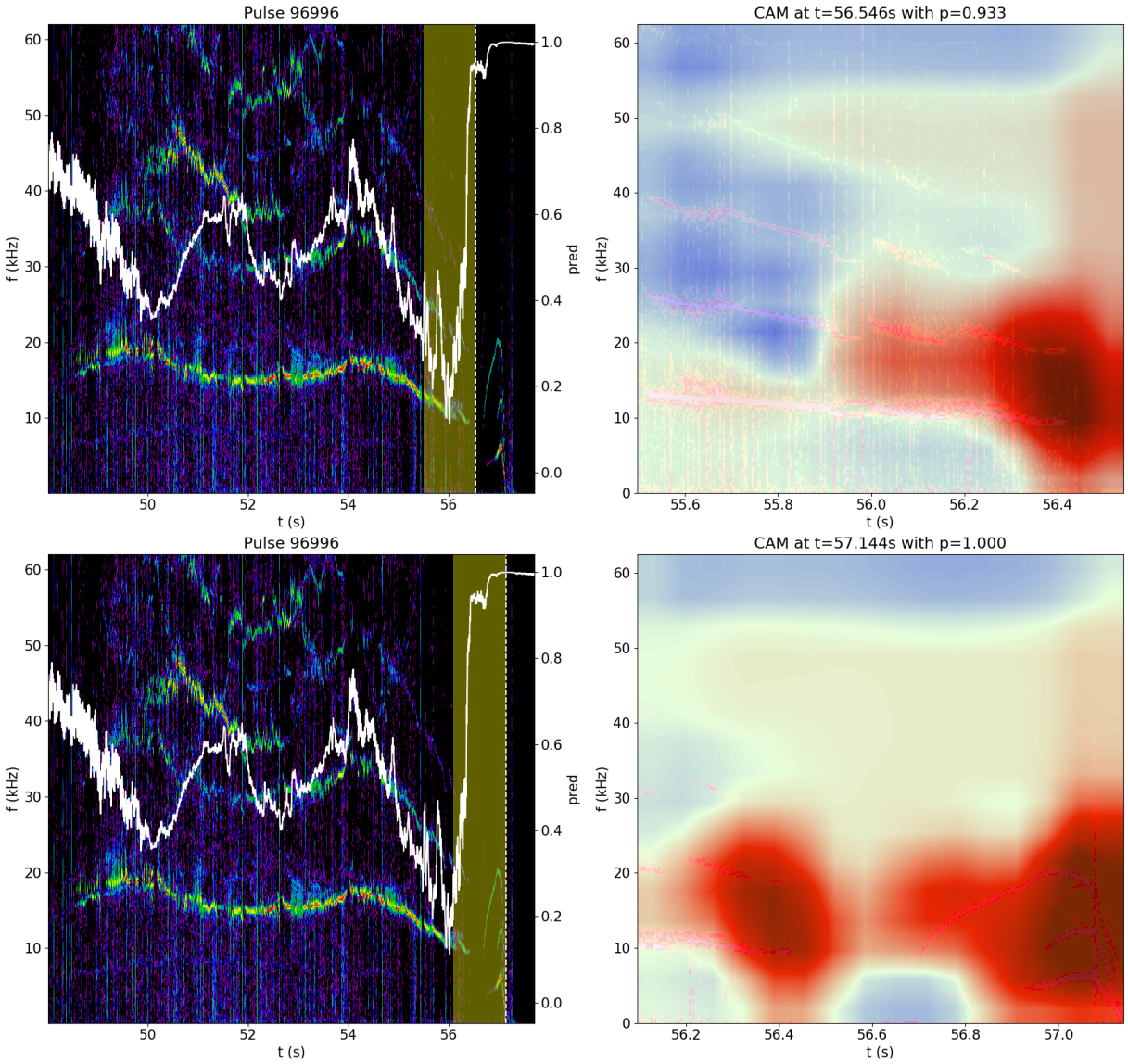


Figure 6.10: Result of CAM for discharge 96996 at t=56.54 s (top) and t=57.14 s (bottom).

Again, the considerable time difference between the alarm provided with the proposed method and the B_{LM} threshold (figure 6.11) is an interesting observation to be retained.

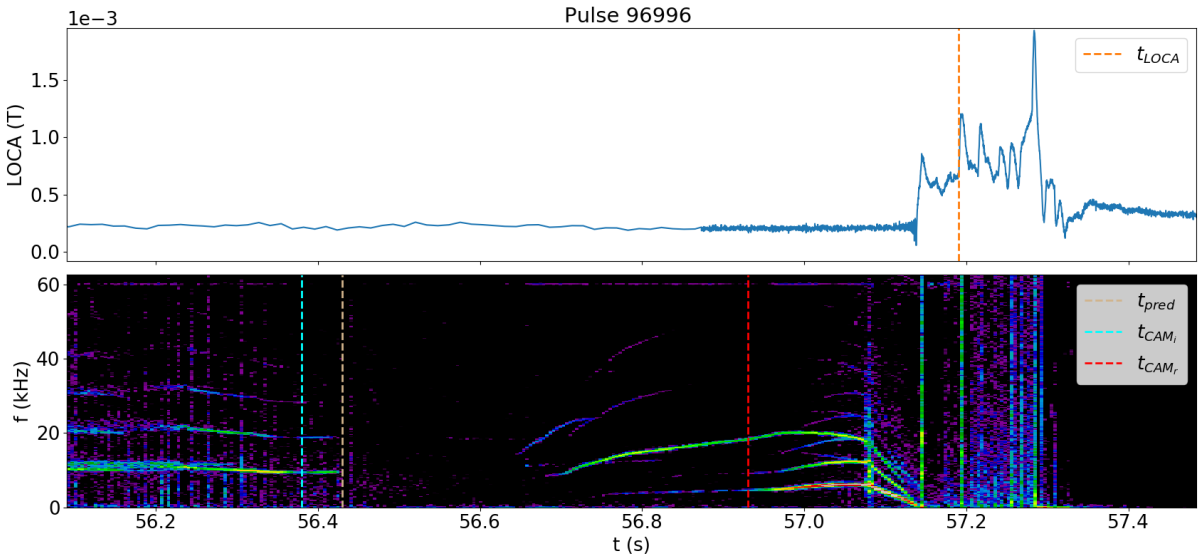


Figure 6.11: Locked mode threshold (top), CAM highlights and probability threshold (bottom) at discharge 96996.

6.5 Discussion

The results obtained from the model's performance metrics tell us that, despite the possible limitations regarding the input type and the training method, they are quite reasonable as a first approach, having into account that just one feature (i.e., the H305 signal) is being given to the model and that no constraints are considered when selecting spectrogram samples. The apparent noise in the prediction value is also a limitation which raises due to the nature of the input. These limitations do not yet allow a direct comparison with other approaches, although this was not the main goal of this work.

Regarding the application of CAM, it is interesting to see how the model was able to agree with the physical interpretation already established for similar disruptive discharges, even though no additional information about the experiment was provided to our model.

The CAM method was applied to practically all discharges in the dataset. In 22 discharges (excluding discharges 92213 and 96996), the same pattern described previously is explicitly observed, that is, the interruption of the MHD activity and its resurgence, followed by the locked mode. All these discharges surpass the prediction threshold of the model, thus being classified as disruptive, which is consistent with the prior attributed labels. In some cases, either the prediction value or the positive highlight of CAM is before mode-locking and could provide a sooner warning time than a simple threshold on the locked mode amplitude, as seen in figures 6.9 and 6.11.

The mentioned discharges can be seen in table 6.3, where the alarm provided by the CNN classifier, t_{pred} is also compared with the time when the B_{LM} threshold is surpassed, defined as t_{LOCA}. The difference is measured with $\Delta t = t_{LOCA} - t_{pred}$. Our approach could improve significantly the warning

Discharge	$t_{LOCA}(s)$	$t_{pred}(s)$	$\Delta t(s)$
89762	53.41	47.55	5.86
89763	53.31	47.55	5.76
90319	51.57	47.55	4.02
90322	48.53	47.16	1.37
91971	51.89	48.61	3.28
91972	50.22	49.86	0.36
92215	53.59	52.04	1.55
92424	53.37	51.65	1.72
92427	51.58	51.16	0.42
92432	52.37	50.81	1.56
92433	52.38	50.13	2.25
92435	53.99	52.81	1.18
94423	53.34	48.42	4.92
94903	55.95	53.81	2.14
95531	52.35	52.12	0.23
95541	55.56	49.98	5.58
95997	53.72	52.11	1.61
96468	52.03	51.72	0.31
96486	55.47	51.82	3.65
96537	53.67	52.27	1.40
96885	56.81	52.97	3.84
96892	56.83	55.40	1.43

Table 6.3: Identified discharges with the CAM highlighting for the interruption and resurgence of MHD activity.

in the mentioned discharges. Having a larger time window due to higher warning times can be relevant when applying these methods in real-time applications to support the DMS in a tokamak.

However, these conclusions need to be carefully constrained. In some of the identified discharges, the CAM method also gives a particular focus to the described patterns, but it may not increase considerably the probability value in those windows, as previous MHD activity could enhance it. Additionally, the presence of many FP discharges when the calculations of the performance metrics were made can indicate that the instability of the probability values throughout each discharge makes the model wrongly classify some of the non-disruptive discharges as disruptive. To conclude, some discharges could be left out of the CAM behavior identification due to the short duration of the locked mode.

7

Conclusions

Contents

In this work, we have shown a deep learning method to be applied directly on spectrogram data with MHD activity to predict the occurrence of disruptions with mode-locking. Although it cannot be directly compared with other predictors, it can reasonably discriminate the classes of interest, having into account that only the data coming from a magnetic coil was used as an input. Additionally, the CAM method was implemented to allow the addition of physical understanding of the results and retrieve relevant MHD activity in the perspective of the model.

Based on the application of CAM mapping to various discharges, the most common feature to be compatible with available physical insight is the interruption of MHD activity, followed by its resurgence. In some discharges, the model can provide an alarm before mode-locking itself, and it can even surpass current predictors in the time window for prediction.

To sum up, this work delivers a new technique to analyze MHD activity in time and frequency representations, as well as provides some insight when data-driven models are applied in nuclear fusion data. Furthermore, it contributes to the new paradigm in the disruption prediction field, where besides the capabilities of the predictors to correctly classify disruptive discharges, it allows to increase the knowledge in the mechanisms involved, as well as to validate the results from the applied deep learning models in fusion data. The introduction of this methodology and other interpretable deep learning frameworks can be an important analysis tool for tokamak physicists, allowing a broader perspective of data-driven models and providing the learned features from the models to extend the analysis of MHD activity and

other machine diagnostics.

7.1 Contributions

The main contributions of this work can be listed as follows:

- According to the analyzed dataset of JET discharges from the baseline scenario, there is an intrinsic correlation between the locked mode and disruptive discharges. A threshold value of its amplitude can clearly separate non-disruptive and disruptive classes. This is consistent with the extensive bibliography on the subject.
- A CNN model was developed to predict the occurrence of the locked mode from spectrogram samples of a given discharge. Although it is not directly comparable with other predictors, the results suggest that it can be useful as a disruption predictor that anticipates the occurrence of the locked mode itself.
- The warning time provided with the described approach in some cases is significantly higher when compared with the locked mode amplitude and other methods. Our model, for instance, could contribute to the disruption mitigation field and be incorporated with other systems and predictors of current and future tokamaks.
- The CAM technique was tested for the first time, at least to our knowledge, in data from magnetic coils. This provided some interpretability to the results obtained from the CNN and thus an interconnection with the physical analysis could be made.
- Our results suggest that the most common behavior obtained by CAM was the interruption and resurgence of MHD activity, followed by mode-locking. This is consistent with observed behavior in other studies. It is also a demonstration of congruence between what the model retrieves and the physics point of view.

Still, certain limitations should be mentioned. As stated previously in this work, there are additional mechanisms in tokamak disruptions apart from the locked mode. To build a predictor for a broader range of disruptive mechanisms, it is necessary to take into consideration other diagnostics from the machine. This inclusion could also bring some improvements to our model when combined with the spectrogram data.

7.2 Future work

A new set of possible additional tasks arise after the first approach with these methodologies:

- Parameter and hyperparameter optimization of the model. A manual search of both parameters and hyperparameters was done to reach an optimal combination of performance metrics of our model. These parameters spaces were not completely explored, and a more extensive tuning, with proper computational resources, could lead to improvements in their accuracy.
- Extension of the methodology to more experimental scenarios of JET. Since the dataset was considerably limited in the number of discharges, the inclusion of more experiments with different experimental conditions could also improve the model performance and bring additional insight.
- Establishment of a new type of labeling and classification constraints. Using binary labels can be a limitation, specifically when the samples are at the beginning of the discharge. The time from which these samples are considered can as well be delimited by other plasma diagnostics, for example.
- The addition of other interpretability techniques. In the implementation of CAM, it is necessary to change the original model to obtain the weighted sum of feature maps. There are other methods where this is not necessary that can be applied to more complex models.
- The examination of other behaviors retrieved by CAM. Besides the mentioned marker on the interruption and resurgence of MHD activity, a more detailed analysis of the method throughout all discharges could bring additional insight.

Bibliography

[1] A. S. Eddington, “The Internal Constitution of the Stars,” *Nature*, vol. 106, pp. 14–20, 1920.

[2] H. A. Bethe, “Energy Production in Stars,” *Phys. Rev.*, vol. 55, pp. 434–456, 1939.

[3] J. P. Freidberg, *Plasma Physics and Fusion Energy*, 1st ed. Cambridge University Press, 2008.

[4] A. Vallet, “Hydrodynamic modelling of the shock ignition scheme for inertial confinement fusion (’Modélisation hydrodynamique du schéma d’allumage par choc pour la fusion par confinement inertiel’).” Ph.D. dissertation, L’Université de Bordeaux, 2014.

[5] F. F. Chen, *Introduction to Plasma Physics and Controlled Fusion*, 3rd ed. Springer, 2016.

[6] C. Hopf, G. Starnella, N. den Harder and U. Fantz, “Neutral beam injection for fusion reactors: technological constraints versus functional requirements,” *Nucl. Fusion*, vol. 61, p. 106032, 2021.

[7] Ye. O. Kazakov et al., “Efficient generation of energetic ions in multi-ion plasmas by radio-frequency heating,” *Nature Physics*, vol. 13, p. 973–978, 2017.

[8] J. Wesson and D.J. Campbell, *Tokamaks*, 1st ed. Clarendon Press - Oxford, 2004.

[9] National Research Council, *Burning Plasma: Bringing a Star to Earth*, 1st ed. The National Academies Press, 2004.

[10] T. C. Hender et al., “Chapter 3: MHD stability, operational limits and disruptions,” *Nucl. Fusion*, vol. 47, pp. S128–S202, 1999.

[11] A. H. Boozer, “Theory of tokamak disruptions,” *Physics of Plasmas*, vol. 19, no. 5, p. 058101, 2012.

[12] P. De Vries et al., “Requirements for triggering the iter disruption mitigation system,” *Fusion Science and Technology*, vol. 69, p. 471–484, 2016.

[13] P. C. De Vries et al., “Survey of disruption causes at JET,” *Nucl. Fusion*, vol. 51, p. 053018, 2011.

[14] G. H. Neilson, *Magnetic Fusion Energy*, 1st ed. Woodhead Publishing, 2016.

[15] M.Lehen et al., “Disruptions in ITER and strategies for their control and mitigation,” *Journal of Nuclear Materials*, vol. 463, pp. 39–48, 2015.

[16] A. Hassanein and V. Sizyuk, “Potential design problems for ITER fusion device,” *Sci. Rep.*, vol. 11, p. 2069, 2021.

[17] J. Kates-Harbeck, A. Svyatkovskiy, and W. Tang, “Predicting disruptive instabilities in controlled fusion plasmas through deep learning,” *Nature*, vol. 568, pp. 526–531, 2019.

[18] J. Vega et al., “Disruption precursor detection: Combining the time and frequency domains,” in *2015 IEEE 26th Symposium on Fusion Engineering*, 2015, pp. 1–8.

[19] D. R. Ferreira, “Applications of deep learning to nuclear fusion research,” 2018, arXiv:1811.00333.

[20] C.E. Kessel et al., “Development of ITER 15 MA ELMy H-mode inductive scenario,” *Nucl. Fusion*, vol. 49, pp. 85 034–85 053, 2009.

[21] E.J. Strait et al., “Progress in disruption prevention for ITER,” *Nucl. Fusion*, vol. 59, p. 112012, 2019.

[22] C. Sozzi et al., “Early identification of disruption paths for prevention and avoidance,” in *27th IAEA Fusion Energy Conference*, 2018, pp. 216–217.

[23] F. A. Volpe et al., “Avoiding tokamak disruptions by applying static magnetic fields that align locked modes with stabilizing wave-driven currents,” *Phys. Rev. Lett.*, vol. 115, p. 175002, 2015.

[24] J. A. Wesson et al., “Disruptions in JET,” *Nucl. Fusion*, vol. 29, p. 066028, 1989.

[25] H. Dreicer, “Electron and ion runaway in a fully ionized gas. i,” *Phys. Rev.*, vol. 115, pp. 238–249, 1959.

[26] J. W. Connor and R.J. Hastie, “Relativistic limitations on runaway electrons,” *Nucl. Fusion*, vol. 15, pp. 415–424, 1975.

[27] A. Loarte et al., “Magnetic energy flows during the current quench and termination of disruptions with runaway current plateau formation in JET and implications for ITER,” *Nucl. Fusion*, vol. 51, p. 073004, 2011.

[28] J.R. Martín, A. Loarte and M. Lehnen, “Formation and termination of runaway beams in ITER disruptions,” *Nucl. Fusion*, vol. 57, p. 066025, 2017.

[29] F C Schuller, “Disruptions in tokamaks,” *Plasma Phys. and Control. Fusion*, vol. 37, pp. A135–A162, 1995.

[30] M. Murakami, J.D. Callen and L.A. Berry, “Some observations on maximum densities in tokamak experiments,” *Nucl. Fusion*, vol. 16, pp. 347–348, 1976.

[31] M. Greenwald et al., “A new look at density limits in tokamaks,” *Nucl. Fusion*, vol. 28, no. 12, pp. 2199–2207, 1988.

[32] J. A. Wesson, “Hydromagnetic stability of Tokamaks,” *Nucl. Fusion*, vol. 18, pp. 87–132, 1978.

[33] F. Troyon et al., “MHD-Limits to Plasma Confinement,” *Plasma Phys. and Control. Fusion*, vol. 26, pp. 209–215, 1984.

[34] R. G. Kleva and P. N. Guzdar, “Fast disruptions by ballooning mode ridges and fingers in high temperature, low resistivity toroidal plasmas,” *Physics of Plasmas*, vol. 8, no. 1, p. 103–109, 2001.

[35] M. Ariola, G. De Tommasi, A. Pironti, and F. Villone, “Control of resistive wall modes in tokamak plasmas,” *Control Engineering Practice*, vol. 24, pp. 15–24, 2014.

[36] D. D. Schnack, “Ideal MHD and the Frozen Flux Theorem,” June 1984.

[37] D. Batchelor et al., “Simulation of wave interactions with MHD,” *Journal of Physics: Conference Series*, vol. 125, p. 012039, 2008.

[38] H. P. Furth, J. Killeen, and M. N. Rosenbluth, “Finite Resistivity Instabilities of a Sheet Pinch,” *Phys. Fluids*, vol. 6, pp. 459–484, 1963.

[39] Z. Chang and J.D. Callen, “Global energy confinement degradation due to macroscopic phenomena in tokamaks,” *Nucl. Fusion*, vol. 30, pp. 219–233, 1990.

[40] O. Sauter et al., “Beta limits in long-pulse tokamak discharges,” *Physics of Plasmas*, vol. 4, no. 5, pp. 1654–1664, 1997.

[41] C. C. Hegna, “The physics of neoclassical magnetohydrodynamic tearing modes,” *Physics of Plasmas*, vol. 5, no. 5, pp. 1767–1774, 1998.

[42] M. Baruzzo et al., “Neoclassical tearing mode (NTM) magnetic spectrum and magnetic coupling in JET tokamak,” *Plasma Phys. and Control. Fusion*, vol. 52, p. 075001, 2010.

[43] R. Fitzpatrick, “Interaction of tearing modes with external structures in cylindrical geometry (plasma),” *Nucl. Fusion*, vol. 33, pp. 1049–1084, 1993.

[44] R. Sweeney et al., “Statistical analysis of m/n = 2/1 locked and quasi-stationary modes with rotating precursors at DIII-D,” *Nucl. Fusion*, vol. 57, p. 016019, 2017.

[45] M.F. Nave and J.A. Wesson, “Mode locking in tokamaks,” *Nucl. Fusion*, vol. 30, pp. 2575–2583, 1990.

[46] S. N. Gerasimov et al., “Locked mode and disruptions in JET-ILW,” in *46th EPS Conference on Plasma Physics*, 2019.

[47] H. Strauss, “Thermal quench in ITER locked mode disruptions,” *Physics of Plasmas*, vol. 28, no. 7, p. 072507, 2021.

[48] C. Reaux et al., “Use of the disruption mitigation valve in closed loop for routine protection at JET,” *Fusion Engineering and Design*, vol. 88, p. 1101, 2013.

[49] U. Kruezi et al., “Massive gas injection experiments at JET - performance and characterisation of the disruption mitigation valve,” 2009.

[50] P. C. De Vries et al., “Scaling of the MHD perturbation amplitude required to trigger a disruption and predictions for ITER,” *Nucl. Fusion*, vol. 56, p. 026007, 2015.

[51] L. R. Baylor et al., “Shattered pellet injection technology design and characterization for disruption mitigation experiments,” *Nucl. Fusion*, vol. 59, p. 066008, 2019.

[52] C. Rea et al., “Disruption prediction investigations using Machine Learning tools on DIII-D and Alcator C-Mod,” *Plasma Phys. and Control. Fusion*, vol. 60, p. 084004, 2018.

[53] B. Cannas et al., “Support vector machines for disruption prediction and novelty detection at JET,” *Fusion Engineering and Design*, vol. 82, pp. 1124–1130, 2007.

[54] J. M. Lopez et al., “Implementation of the disruption predictor APODIS in JET’s real-time network using the MARTe framework,” in *IEEE Transactions on Nuclear Science*, 2012, pp. 1–4.

[55] J Croonen, J. Amaya, and G. Lapenta, “Tokamak disruption prediction using different machine learning techniques,” 2020, arXiv:2005.05139.

[56] B. Cannas et al., “A prediction tool for real-time application in the disruption protection system at JET,” *Nucl. Fusion*, vol. 47, pp. 1559–1569, 2007.

[57] J. Vega et al., “Results of the JET real-time disruption predictor in the ITER-like wall campaigns,” *Fusion Engineering and Design*, vol. 88, pp. 1228–1231, 2013.

[58] R. M. Churchill et al., “Deep convolutional neural networks for multi- scale time-series classification and application to tokamak disruption prediction using raw, high temporal resolution diagnostic data,” *Physics of Plasmas*, vol. 27, no. 6, p. 062510, 2020.

[59] A. Svyatkovskiy, J. Kates-Harbeck, and W. Tang, “Training distributed deep recurrent neural networks with mixed precision on GPU clusters,” *Proceedings of the Machine Learning on HPC Environments*, 2017.

[60] A. Agarwal et al., “Using LSTM for the Prediction of Disruption in ADITYA Tokamak,” 2020, arXiv:2007.06230.

[61] B. Cannas et al., “Automatic disruption classification in JET with the ITER-like wall,” *Plasma Phys. and Control. Fusion*, vol. 57, p. 125003, 2015.

[62] J. Vega et al., “A linear equation based on signal increments to predict disruptive behaviours and the time to disruption on JET,” *Nucl. Fusion*, vol. 60, p. 026001, 2020.

[63] A. Murari et al., “Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions,” *Nucl. Fusion*, vol. 53, p. 033006, 2013.

[64] D. R. Ferreira et al., “Deep Learning for the Analysis of Disruption Precursors Based on Plasma Tomography,” *Fusion Science and Technology*, vol. 76, pp. 901–911, 2020.

[65] A. Murari et al., “Investigating the physics of tokamak global stability with interpretable machine learning tools,” *Applied Sciences*, vol. 10, no. 19, 2020.

[66] C. Rea et al., “Progress Toward Interpretable Machine Learning–Based Disruption Predictors Across Tokamaks,” *Fusion Science and Technology*, vol. 76, pp. 912–924, 2020.

[67] L. H. Gilpin et al., “Explaining explanations: An overview of interpretability of machine learning,” 2019, arXiv:1806.00069.

[68] B. Zhou et al., “Learning Deep Features for Discriminative Localization,” 2015, arXiv:1512.04150.

[69] G. Kwon, H. Wi, and J. Hong, “Tokamak visible image sequence recognition using nonlocal spatio-temporal CNN for attention needed area localization,” *Fusion Engineering and Design*, vol. 168, p. 112375, 2021.

[70] G. Sias et. al., “A locked mode indicator for disruption prediction on JET and ASDEX upgrade,” *Fusion Engineering and Design*, vol. 138, pp. 254–266, 2019.

[71] S. Esquembri et al., “Real-Time Implementation in JET of the SPAD Disruption Predictor Using MARTe,” *IEEE Transactions on Nuclear Science*, vol. 65, pp. 836–842, 2018.

[72] J. V. Hernandez et al., “Neural network prediction of some classes of tokamak disruptions,” *Nucl. Fusion*, vol. 36, pp. 1009–1017, 1996.

[73] A. Bustos, E. Ascasíbar, A. Cappa, and R. Mayo-García, “Automatic Identification of MHD Modes in Magnetic Fluctuations Spectrograms using Deep Learning Techniques,” *Plasma Phys. and Control. Fusion*, vol. 63, p. 095001, 2021.

[74] L. Garzotti et al., “Scenario development for D–T operation at JET,” *Nucl. Fusion*, vol. 59, p. 076037, 2019.

[75] The European Commission, *THE JET PROJECT: Design Proposal for the Joint European Torus*. The European Commission, 2008.

[76] M. Keilhacker, A. Gibson, C. Gormezano and P.H. Rebut, "The scientific success of JET," *Nucl. Fusion*, vol. 41, pp. 1925–1966, 2001.

[77] T. Fujita et al., "High performance experiments in JT-60U reversed shear discharges," *Nucl. Fusion*, vol. 39, p. 1627, 2002.

[78] J. Wesson, *The Science of JET*, 2nd ed. JET Joint Undertaking, 2006.

[79] I. H. Hutchinson, *Principles of Plasma Diagnostics*, 2nd ed. Cambridge University Press, 2002.

[80] M. F. F. Nave et al., "On the use of MHD mode analysis as a technique for determination of q-profiles in JET plasmas," *Review of Scientific Instruments*, vol. 75, no. 10, pp. 4274–4277, 2004.

[81] G. Artaserse et. al., "Refurbishment of JET magnetic diagnostics," *Fusion Engineering and Design*, vol. 146, pp. 2781–2785, 2019.

[82] H. Jeon, Y. Jung, S. Lee, and Y. Jung, "Area-Efficient Short-Time Fourier Transform Processor for Time–Frequency Analysis of Non-Stationary Signals," *Applied Sciences*, vol. 10, no. 20, 2020.

[83] J. P. Bizarro and A. Figueiredo, "Time–frequency analysis of fusion plasma signals beyond the short-time Fourier transform paradigm: An overview," *Fusion Engineering and Design*, vol. 83, pp. 350–353, 2008.

[84] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 6, 2020.

[85] J. Madrid-Sánchez, M. Lázaro-Gredilla, and A. R. Figueiras-Vidal, "A Single Layer Perceptron Approach to Selective Multi-task Learning," in *Bio-inspired Modeling of Cognitive Tasks*, 2007, pp. 272–281.

[86] I. Gómez, L. Franco, J. Subirats, and J. Jerez, "Neural network architecture selection: Size depends on function complexity," in *ICANN*, 2006, pp. 122–129.

[87] M. Browne and S. Ghidary, "Convolutional neural networks for image processing: An application in robot vision," in *Australian Conference on Artificial Intelligence*, 2003, pp. 641–652.

[88] M. Dörfler, R. Bammer, and T. Grill, "Inside the spectrogram: Convolutional neural networks in audio processing," in *2017 International Conference on Sampling Theory and Applications*, 2017, pp. 152–155.

[89] D. Carvalho, “Plasma Tomography with Machine Learning,” Master’s thesis, Instituto Superior Técnico, 2018.

[90] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2015.

[91] H. Ide and T. Kurita, “Improvement of learning for CNN with ReLU activation by sparse regularization,” in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2684–2691.

[92] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” 2015, arXiv:1412.6980.

[93] R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[94] S. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, “Impact of fully connected layers on performance of convolutional neural networks for image classification,” *Neurocomputing*, vol. 378, p. 112–119, 2020.

[95] K. Jayech, “Regularized Deep Convolutional Neural Networks for Feature Extraction and Classification,” *Lecture Notes in Computer Science*, vol. 10635, pp. 431–439, 2017.

[96] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[97] G. Pucella et al., “Onset of tearing modes in plasma termination on JET: the role of temperature hollowing and edge cooling,” *Nucl. Fusion*, vol. 61, p. 046020, 2021.