Data-driven study of major disruption prediction and plasma instabilities across multiple tokamaks

by

Jinxiang Zhu

B.S., Zhejiang University (2018)

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2023

Author .....
Department of Physics
Jan 25, 2023

Certified by .....
Earl Marmar
Senior Research Scientist, Department of Physics
Thesis Supervisor

Accepted by .....
Lindley Winslow
Associate Department Head of Physics

Data-driven study of major disruption prediction and plasma instabilities across multiple tokamaks

by

Jinxiang Zhu

Submitted to the Department of Physics
on Jan 25, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Physics

Abstract

The use of nuclear fusion energy via magnetic-confinement tokamaks is one of a few encouraging paths toward future sustainable energy. Along the way, scientists need to learn to avoid plasma disruptions: these sudden and unexpected plasma terminations still represent one of the key challenges for tokamak devices. Forecasting plasma instabilities and disruptions using first-principle models has been demonstrated to be extremely difficult, due to the complexity of the problem and the high non-linearity of the system. To date, disruption and plasma instabilities prediction has been studied through two main approaches: data-driven versus physics-driven (or model-based). On the one hand, recent statistical and machine learning (ML) approaches based on experimental data have shown attractive results for disruption prediction, even in real-time environments. Different tokamak devices have different operational spaces, spatiotemporal scales for physics events, and plasma diagnostics. Therefore, most of these data-driven approaches were developed and optimized specifically for one device and did not show promising cross-device predictive ability. In addition, the complexity of these data-driven models limits their physics interpretability. Recent Deep-Learning (DL) based disruption prediction studies demonstrate the potential for acquiring a general representation of experimental data that can be used in cross-machine applications. On the other hand, model-based studies seek to identify event chains that can lead to disruptions through early event detection, which can help operators to avoid plasma instabilities disruptions. However, the extrapolation ability of physics-based models to new devices, especially to new physics regimes is still unclear.

This thesis demonstrates the application of data-driven methods on plasma instabilities and disruption prediction via four major contributions. First, through explorative data analysis of thousands of shots on C-Mod, DIII-D and EAST tokamaks, the advantage of sequence-based disruption prediction model was shown. Based on this finding, a new Hybrid Deep-Learning (HDL) general disruption predictor was developed using C-Mod, DIII-D and EAST databases and it achieves state-of-the-art performance on three machines with only limited hyperparameter tuning. Dedicated cross-machine disruption prediction studies using this HDL model demonstrated that a significantly boosted accuracy on the target machine was achieved by training on

20 disruptive shots, thousands of non-disruptive shots from the target machine combined with hundreds of disruptive shots from other devices. In addition, by comparing the predictive performance of each individual numerical experiment, the disruptive shots from multiple devices were found to contain device-independent knowledge that can be used to inform predictions for disruptions occurring in a new device while non-disruptive shots were found to be machine-specific. Second, the cross-regime disruption prediction on multiple tokamaks using HDL model demonstrated data-driven disruption predictors trained on abundant Low Performance (LP) discharges work poorly on the High Performance (HP) regime of the same tokamak, which is a consequence of the distinct distributions of the tightly correlated signals related to disruptions in these two regimes. Moreover, the cross machine experiments suggested matching operational parameters among tokamaks strongly improves cross-machine accuracy. Given these conclusions, a scenario adaptive strategy that works for all data-driven models was proposed for next generation tokamaks, such as ITER and SPARC, and highlight the importance of developing baseline scenario discharges of future tokamaks on existing machines to collect more relevant disruptive data. Third, the powerful HDL model was upgraded to an integrated ML model that can predict major disruption as well as multiple unstable events in tokamak plasmas that can facilitate the physics interpretation of output from the black box data-driven models and enables disruption avoidance by responding to early unstable events of plasmas. Enhanced cross-machine ability and improved warning time was also observed using the integrated ML model. Finally, among all different plasma unstable events, the $n = 1$ tearing mode (TM) is considered to be one of the most important disruption precursors and its predictive ability is strongly desirable for ITER and SPARC. In the final part of this thesis, an empirical boundary for the $n = 1$ tearing mode (TM) is developed via data-driven methods and verified on thousands of DIII-D discharges. The fitted boundary is a linear function of plasma equilibrium parameters such as collisionality, poloidal beta, and the MHD risk factor (a combination of the normalized electron temperature profile width, q95 and elongation). The boundary indicates with a value related to the probability of having the TM onset and it achieves 88% of shot-by-shot accuracy in offline analysis of DIII-D data. Preliminary cross-machine analysis of TM onset prediction shows potential applicability of the empirical boundary to C-Mod and EAST data as well, but the relative importance of the individual parameters is different for different devices. This suggests the existence of different trigger mechanisms for the TMs, implying that the boundary could be generalized using data from different tokamaks representing different trigger mechanisms to improve its extrapolability. Finally, this new proximity metric to the $n = 1$ TM onset has been incorporated into the real-time in DIII-D plasma control system (PCS) and results from real-time experiments will be discussed.

Thesis Supervisor: Earl Marmar
Title: Senior Research Scientist, Department of Physics

Acknowledgments

The completion of this thesis and my Ph.D research were made possible by the guidance and support of many people. Among these important people, I would first like to thank my supervisors, Robert Granetz, Cristina Rea and Earl Marmar, their kind instruction gave me invaluable knowledge and expertise and taught me how to gradually transit from a student to a junior researcher. They also provided me lots of resources including collaborators, access to experimental data from various devices, run time on major tokamaks like DIII-D and computational resources. Moreover, Earl met with me regularly to provide support and feedback that improved my work and kept me on track.

Besides my supervisors, I want to thank the community of scientists, staff, and students at the MIT PSFC. Especially, I would like to my thesis committee member and professor, Nuno Loureiro, who spent lots of time giving important feedback of my thesis and allowed me to catch up the schedule for completion in February. I also want to thank the remaining members of the PSFC Disruptions group, including Ryan Sweeney, Alex Tinguely, Kevin Montes, Benjamin Stein-Lubrano, and Andrew Maris, for their supporting work and comments to my research projects included in this thesis.

My research projects included in this thesis is based on experimental data from C-Mod, DIII-D and EAST and the construction of these datasets were made possible via close collaboration with scientists from several research institutions. I am grateful to to my colleagues from the Alcator C-Mod, DIII-D, and EAST teams for their contributions on populating datasets used in this thesis. I owe a special thanks to Kevin Montes whose manually labeled DIII-D dataset gave rise to the research project summarized in chapter 5. I would especially like to recognize Jayson Barr, Ai Hyatt and Tom Osborne from from General Atomics, Francesca Turco from Columbia and Keith Erickson from PPPL for their contributions to the DIII-D TM avoidance experiments planning and for implementing real-time tearing mode boundary on the DIII-D plasma control system.

Finally, I would like to thank my parents, Keming Zhu and Yulan Xie, for raising me and providing me with a great education that allow me to come to MIT and enjoy my Ph.D research at PSFC. I look forward to continuing my research career as a postdoc at PSFC and devoting myself to the SPARC project and commercial fusion.

This research was supported in part by the US Department of Energy, Office of Science, Fusion Energy Sciences, Agreement DE-SC0012071.

Contents

Chapter 1

Introduction

1.1 Nuclear Fusion

There is enormous evidence, including projected near-future energy consumption and the estimation of fossil fuel reserves that point out the need to develop sustainable energy resources [1, 2]. In addition, the burning of fossil fuels is known to cause serious environmental damage and negative impact directly to the earth which, which produces about 90% of all CO_2 emissions [3]. At present, alternative carbon-free energy resources like hydro power, wind power, solar power and bio-energy already play important roles in the world's total energy supplies. However, these technologies are still limited by various unsolved problems including scalability, stability, supply and storage difficulties.

Among all possible future energy resources, the large potential of fusion power makes it hard to be ignored. Fusion energy could provide huge and steady electricity without producing greenhouse gases compared with fossil fuel. In addition, fusion power is expected to have great advantages over current fission power including greatly reduced radioactivity during operation, much less nuclear waste and abundant, widely available fuel resources. Fusion processes occur when two or more nuclei fuse into heavier nuclei. For nuclei lighter than iron-56, the produced heavier nuclei will be slightly lighter than the reactant nuclei. The decrease of mass in the fusion processes, Δm, is converted to kinetic energy of the products according to the mass-energy equation $\Delta E = \Delta m c^2$. In order for the occurrence of nuclear fusion, the reactant nuclei must be given enough kinetic energy to bring them close enough for enough time such that the strong force pulling nuclei together exceeds the electrostatic repulsion. The amount of kinetic energy needed in the process is called the *Coulomb Barrier*. To provide enough kinetic energy, heating the atoms to high temperatures is one of the practical ways. Once the atoms are heated to exceed their ionization energy, they will be ionized to bare nuclei and free electrons. The result is an ionized gas of charged

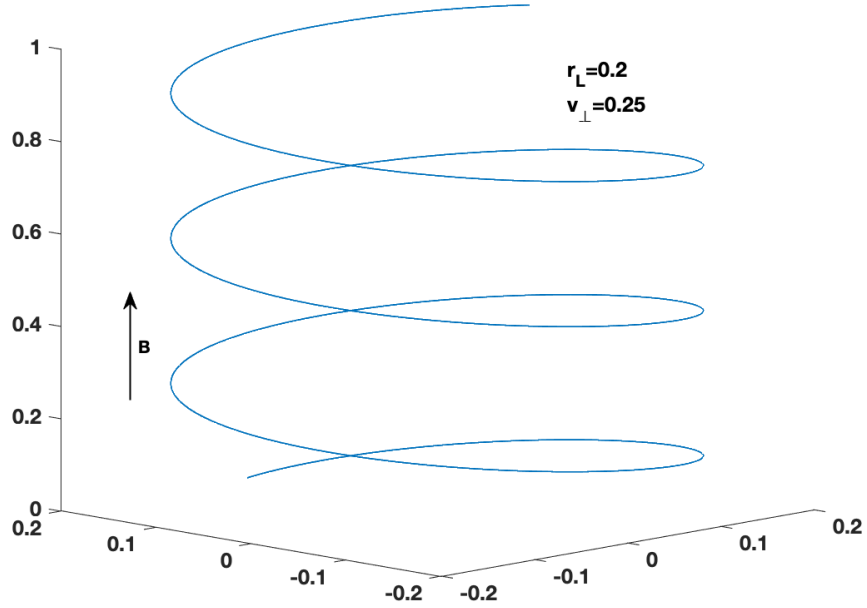

Figure 1-1: The helical orbit of an ion (mass m, charge $q > 0$) moving in a constant magnetic field $\mathbf{B}$ with velocity $v_\perp$ perpendicular to the field and the resulting Larmor radius $r_L = \frac{mv_\perp}{qB}$. This is known as the *Larmor orbit*.

particles known as *plasma*.

Given the reactant nuclei, the probability of having the fusion reaction is described by the reaction *cross section*, σ which depends on the relative velocity of reactant nuclei. In a plasma, particle velocity can be described by a probability distribution. The velocity averaged cross section $\langle \sigma v \rangle$ that depend on the averaged kinetic energy of reactant ions in a plasma is introduced to describe the fusion reaction rate. The highest cross section that can be utilized for fusion energy comes from the reaction of hydrogen isotopes, deuterium-tritium (D-T) reaction when $T_i \sim 15\,\text{keV}$ [4]. The D-T reaction will generate an α particle (He^{2+}) and a neutron (n) with energies $E_\alpha = 3.5\,\text{MeV}$ and $E_n = 14.1\,\text{MeV}$. Since plasma will lose energy through conduction, convection and radiation, to sustain thermonuclear fusion, the sum of input power and the generated fusion power must overcome the loses.

1.2 Magnetic confinement fusion

To sustain the hot plasma cloud, people need to effectively confine the charged particles. After decades of exploration, two current leading confinement schemes are *magnetic confinement* fusion (MCF) using magnetic field and *inertial confinement*

fusion (ICF) by lasers. In MCF, the thermalized plasma is confined by carefully designed magnetic fields. When a charged particle moves in the magnetic fields, its motion is determined by the Lorentz force. For a particle with charge q and velocity v and local magnetic field B, the Lorentz force is $\vec{F} = q\vec{v} \times \vec{B}$. If the magnetic field is uniform, the charged particle will move along the B field with helical path, as shown in Figure 1-1. The gyration radius perpendicular to the field line, known as the *Larmor radius* r_L, is determined by the mass m, charge q and the velocity perpendicular to the field $v_\perp$ of the particle as well as the field strength B with $r_L = \frac{mv_\perp}{qB}$. However, this picture becomes much more complex with the presence of electric field and/or time/space varying magnetic field. These additional complexities usually lead to the *cross-field transport* of the particles by inducing various drift that can cause them leave from the original Larmor orbit and enter the orbit of adjacent field lines. Therefore, the magnetic configuration of a MCF machine must be carefully designed to reduce the transport losses.

For a MCF device sustaining fusion plasmas, the amount of input energy plus generated energy from fusion reaction should overcome various plasma energy losses. Since the amount of energy released in a given volume is a function of reaction rate and hence the ion temperature, the density of reactant ions and the energy confinement time (the length of time that the energy confined in the given volume). Therefore, an important formula, known as the *Lawson criterion* was developed by John D. Lawson in his 1957 paper [5]. Lawson Criterion describes the energy balance for any fusion device based on a hot plasma. In a plasma, a global energy confinement time τ_E measures the system's energy loss rate. It is the internal energy density $W = \frac{3}{2}(n_e T_e + \sum_i n_i T_i)$ divided by power loss density $P_{loss} = P_{input} - dW/dt$ where P_{input} is the total input power density.

$$\tau_E = \frac{W}{P_{input} - dW/dt} \tag{1.1}$$

As mentioned above, the deuterium-tritium (D-T) reaction at $T_i \sim$15 keV gives the highest cross section. For a D-T plasma in the optimum 50-50 mixture and assuming all species have the same temperature and ion density equals electron density, the internal energy density W is given by $W = 3nT$ where n is the particle density and T is the temperature in eV. Given that the produced neutron is electrically neutral, it cannot be confined by the magnetic field and quickly escapes the plasma. Therefore, it does not help to heat the plasma. On the contrary, the α particle is charged and is confined in the plasma by magnetic field. Thus, the α particle can transfer its energy to other particles through Coulomb collisions and heat the plasma. The Lawson criterion requires the fusion heating power ($P_\alpha = n_D n_T \langle \sigma v \rangle E_\alpha$ with n_D, n_T corresponding to the density of D, T respectively) plus any external heating power

exceeds the losses. When the fusion heating P_α is large enough, the external heating is no longer required and the plasma reaches ignition. The ignition condition from the Lawson criterion is given by

$$n_i \tau_E \gtrsim \frac{12 T_i}{\langle \sigma v \rangle E_\alpha} \tag{1.2}$$

Based on the above Equation (1.2), an important figure of merit, fusion *triple product* $n_i T_i \tau_E$, gives the criterion for ignition. For D-T reaction, the minimum required triple product occurs at $T_i \sim$14 keV is about $n_i T_i \tau_E \geq 3 \times 10^{21}$keV m^{-3} [6]. An additional important concept of fusion is called *breakeven*. It is a point at which the fusion power carried by the neutrons $P_{fus} \equiv n_D n_T \langle \sigma v \rangle (E_n + E_\alpha) = P_{input}$. However, no magnetic confinement experiment has achieved breakeven yet.

1.3 Tokamak Plasmas

The *tokamak* is a high-performance magnetic confinement device that confines hot fusion plasma in a torus with strong magnetic field. As mentioned above, the charged particle in magnetic field can freely move along the field line and the time/space varying electromagnetic field can give rise to various drift that allow particles to enter the adjacent field lines. To solve these problems, the tokamak magnetic configuration has two key components: a powerful toroidal magnetic field generated by current in external coils, and a poloidal magnetic field primarily induced by a toroidal current I_p driven in the plasma. The combination of these two components gives a helical field line around the torus as shown by Figure 1-2, which overcomes charge gathering drifts on average.

The tokamak idea was initially proposed by Soviet physicists Igor Tamm and Andrei Sakharov in the 1950s. The first working tokamak was the T-1 tokamak in Russia which started operation in 1958 [7]. Since then, many experimental tokamaks have been built ($\sim$ 35 of them were operating when this thesis was written). Right now, the Joint European Torus (JET), located at Culham Centre for Fusion Energy in Oxfordshire, UK is the world's largest tokamak and it achieved world record fusion energy of 59 MJ for 177 MJ of input heating energy. Next generation tokamaks like ITER [8] and SPARC [9] aiming at exceeding breakeven are currently under design and construction. SPARC is expected to operate in a burning plasma with significant α heating within 10 years.

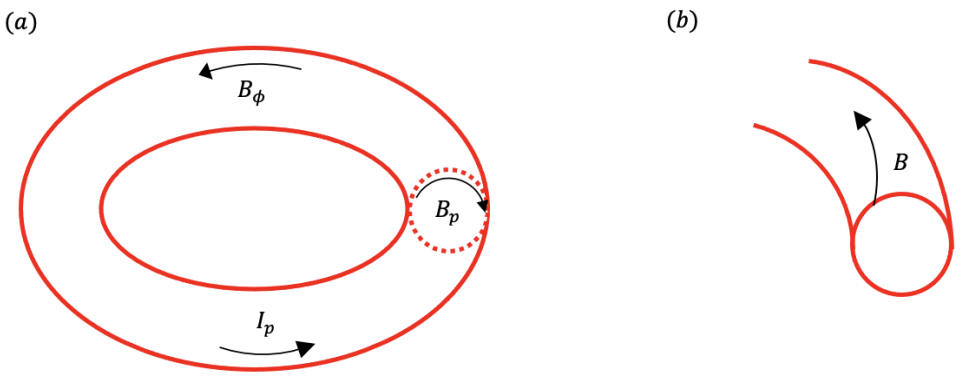

Figure 1-2: (a) Toroidal magnetic field B_ϕ and poloidal magneric field B_p due to toroidal plasma current I_p. (b) Combining B_ϕ and B_p gives helical field lines winding around the torus

1.4 MHD equilibrium of tokamak plasma

Magnetohydrodynamics, or *MHD* is a theory describing the dynamics and electromagnetic properties of an electrically conducting fluid. It applies to large scale (characteristic scale $\gg$ ion Larmor radius and mean free path length) and relatively slow (characteristic time $\gg$ ion gyration time and mean free path time, non-relativistic) plasma which makes it suitable for the global dynamics of tokamak plasmas [10]. The momentum equation from MHD is

$$\rho(\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla)\mathbf{v} = \mathbf{j}\times\mathbf{B} - \nabla p \tag{1.3}$$

where $\mathbf{j}$ is the current density in the plasma and $\mathbf{v}$ is the bulk plasma velocity. From the momentum equation, the equilibrium condition from MHD is the balance of plasma pressure $p = 2nT$ and the magnetic force as

$$\mathbf{j}\times\mathbf{B} = \nabla p \tag{1.4}$$

From Equation (1.4), it is clear that magnetic field lines and currents lie on surfaces of constant p. Therefore, the coordinate transformation can be applied to Equation (1.4) by introducing two flux functions ψ and f [6] that satisfy

$$\mathbf{j}_p = \frac{1}{R}(\nabla f\times\hat{\mathbf{e}}_\phi) \quad \text{and} \quad \mathbf{B}_p = \frac{1}{R}(\nabla\psi\times\hat{\mathbf{e}}_\phi) \tag{1.5}$$

where ϕ and p denote the toroidal and poloidal directions, respectively (see Figure 1-2). In the equation Equation (1.5), ψ is proportional to the poloidal magnetic flux

and f is related to poloidal current. Using these two flux functions ψ and f, Equation (1.4) can be expressed as functions of ψ. Combining MHD equilibrium function with Ampere's law gives the Grad-Shafranov equation

$$R^2\nabla\cdot\left(\frac{\nabla\psi}{R^2}\right) = -\mu_0 R^2\frac{dp(\psi)}{d\psi} - \mu_0^2 f(\psi)\frac{df(\psi)}{d\psi} \tag{1.6}$$

characterizing MHD equilibrium in tokamaks and its solution yields reconstructed pressure, current, and magnetic field profiles [11, 12].

Given an equilibrium tokamak plasma, the ratio of the plasma pressure p to the magnetic pressure $\frac{B^2}{2\mu_0}$, $\beta = \frac{p}{B^2/2\mu_0}$, characterizes its confinement efficiency where $B^2 = B_p^2 + B_\phi^2$. Using large aspect ratio approximation $R \gg a$, $B_\phi \approx B_0$ and the edge poloidal magnetic field $B_p(a)^2 \approx \left(\frac{\mu_0 I_p}{2\pi a}\right)^2 \frac{2}{1+\kappa^2}$, β can be expressed with the toroidal and poloidal components $\beta^{-1} \approx \beta_t^{-1} + \beta_p^{-1}$ where

$$\beta_t \equiv \frac{2\mu_0\langle p\rangle}{B_\phi^2} \approx \frac{2\mu_0\langle p\rangle}{B_0^2} \tag{1.7}$$

$$\beta_p \equiv \frac{2\mu_0\langle p\rangle}{B_p(a)^2} \approx \frac{4\pi^2 a^2(1+\kappa^2)\langle p\rangle}{\mu_0 I_p^2} \tag{1.8}$$

B_0 and κ are the toroidal field on axis and plasma elongation respectively and $\langle\rangle$ represents the volume-averaged quantity over the whole plasma. In addition, $W_{mhd} = \frac{3}{2}\langle p\rangle$ is the the total stored kinetic energy of the plasma.

Another important quantity related to stored plasma magnetic energy is the normalized internal inductance per unit length [10], ℓ_i, defined as

$$\ell_i \equiv \frac{L_i/2\pi R_0}{\mu_0/4\pi} = \frac{2L_i}{\mu_0 R_0} \tag{1.9}$$

where L_i is the plasma internal inductance. In an non-elongated, large aspect ratio plasma, a vertical magnetic field B_V given by

$$B_V = \frac{\mu_0 I_p}{4\pi R_0}\left(\beta_p + \frac{\ell_i - 3}{2} + \ln\frac{8R_0}{a}\right) \tag{1.10}$$

must be applied to sustain the MHD equilibrium [12]. Finally, safety factor q which is highly relevant to MHD stability describes the magnetic topology of the plasma. It can be expressed as

$$q = \frac{1}{2\pi}\oint\frac{1}{R}\frac{B_\phi}{B_p}ds \approx \frac{rB_\phi}{R_0 B_p} \tag{1.11}$$

where the line integral is defined over one poloidal turn of a magnetic surface and the

approximation comes from the non-elongated plasma and large aspect ratio limit [6]. Since $q \propto \frac{r}{B_\theta} \approx \frac{2}{\mu_0 \langle \mathbf{j} \rangle_r}$ where $\langle \mathbf{j} \rangle_r$ is the bulk-averaged toroidal current density inside r and the plasma resistivity $\eta \propto T_e^{-3/2}$ [6], for inductive drive, $\mathbf{j}$ is peaked near the core (hottest) and thus the q profile also reaches minimum at the same region. This is not necessarily for non-inductive drive.

1.5 The Tokamak Disruption

The tokamak *disruption* is a dramatic event that suddenly deteriorates the plasma confinement in an unexpected way. The tokamak disruption is usually initiated from a chain of unstable events (disruption precursors) that lead to the loss of plasma confinement. The loss of confinement has two different consequences. In a minor disruption, the confinement loss is followed by the quick loss of a substantial fraction of kinetic energy and small fraction of electromagnetic energy. The plasma equilibrium gradually recovers after this rapid energy loss. In a major disruption, the confinement loss is unrecoverable and it is followed by the **thermal quench** and then **current quench** in which the plasma quickly loses all its kinetic and electromagnetic energy. A typical example of disruption which occurs during the flattop on DIII-D is shown in Figure 1-3 in which T_e and I_p goes to nearly zero in just a few milliseconds. Disruption is a major challenge for tokamak development because it limits the possible operational region of the tokamak and its deleterious consequences can damage the whole fusion device and prevent the realization of a functioning plasma reactor. Furthermore, the fact that no first principle modeling of plasma disruption exists makes disruption a more serious problem. This section reviews several typical events that can lead to disruptions and the important operational limits related to disruptions.

From decades of tokamak research, several identified unstable events and operational limits are found to be highly related to disruptions. Operational limits for steady state are imposed by density limit, pressure limit and q limit (low-q disruptions) [6]. These limits involve a series of different precursors to the final loss of control. Then the loss of plasma control is always occurring consistently with the thermal and current quench. For example, when q_{95} goes to roughly 2 (low-q limit), $n = 1$ or $n = 2$ tearing modes can grow to large amplitude. The large rotating mode will then be decelerated by wall torque and get locked. The large locked mode can lead to the final loss of control and then the major disruptions happen. Investigating these operational limit and some relevant disruption precursor can help us understand typical disruptive chain of events. We can then use these information to design better disruption prediction/avoidance algorithms. These limits and several typical disruption precursors are reviewed in following subsections.

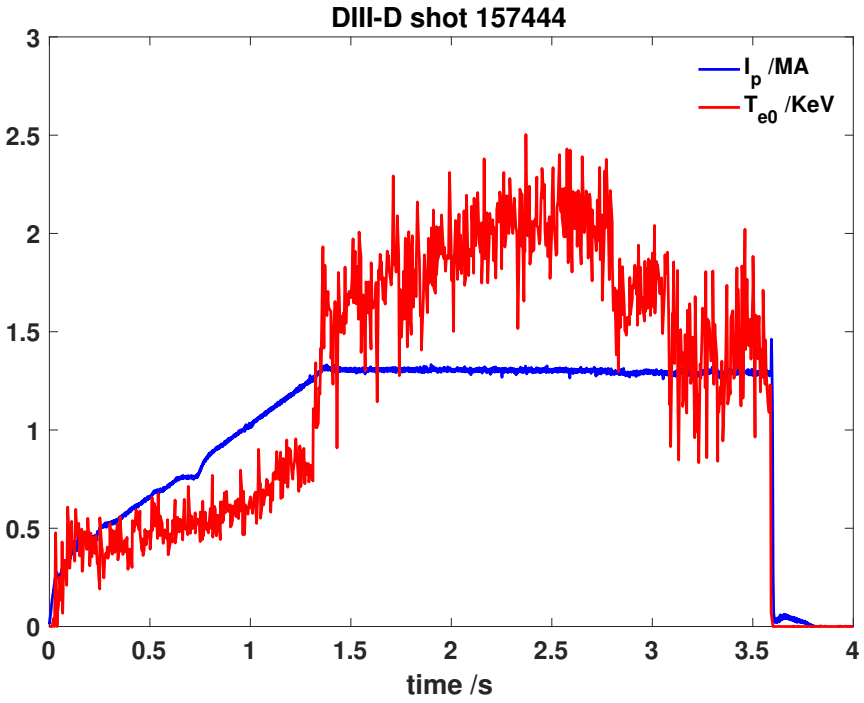

Figure 1-3: Plasma current I_p and central electron temperature T_{e0} for a typical disruption on DIII-D.

1.5.1 Density limit disruption

Observations from various tokamak experiments suggest the maximum density that can be achieved on tokamaks is limited by the scaled plasma current density. This phenomenology was first described by Murakami by empirical scaling [13]. Later, the Greenwald density limit, n_G, was proposed based on experimental data from various machines, including elongated plasmas [14]

$$n_e \lesssim n_G \equiv \frac{I_p}{\pi a^2} \tag{1.12}$$

where n_e is the line averaged electron density with unit 10^{20} m^{-3}, I_p is the plasma current in MA and a is the minor radius in m. A plot shows this n_e vs. n_G scaling is shown in Figure 1-4. The underlying physics of the density limit has not been fully understood yet. Perhaps the most important physics of the density limit is related to the strong onset of particle transport due to increased turbulence at the edge just as the limit is approached. Then, to get higher density through gas fueling requires a dramatic non-linear increase in fueling rate, which leads to the power balance deficit. The increasing density can lead to the increase in impurity radiation and the balance between Ohmic heating ($\propto I_p^2$) and radiation loss from impurities in

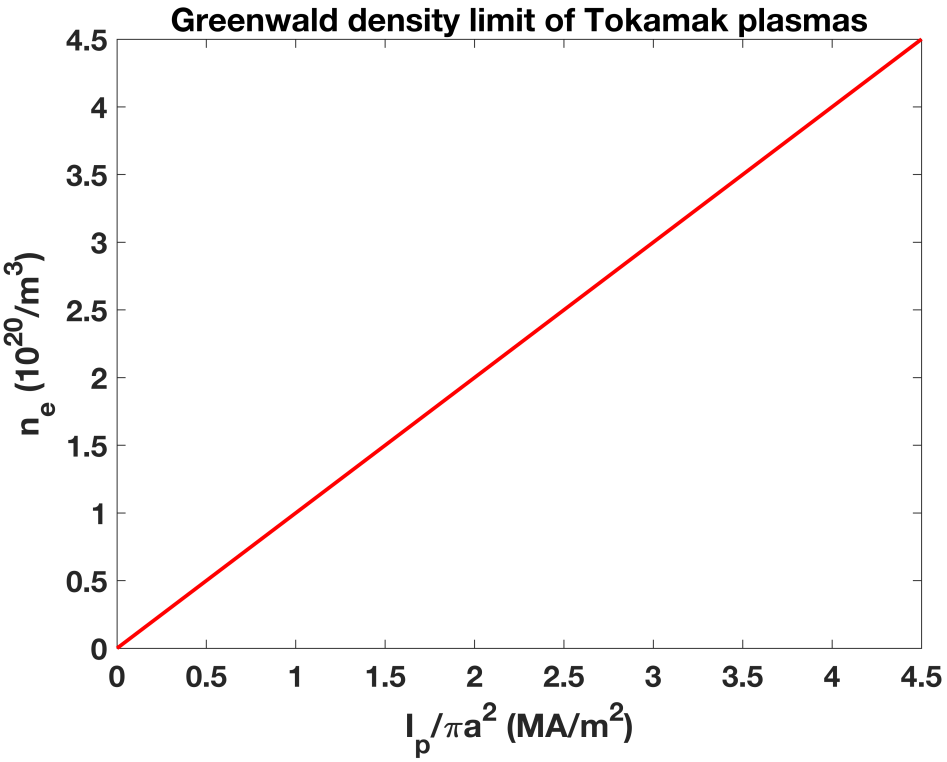

Figure 1-4: Greenwald density limit of Tokamak plasmas.

the cool plasma edge might determine the onset of density limit [15]. Furthermore, the edge cooling usually causes MARFEs (**M**ultifaceted **A**symmetric **R**adiation **F**rom the **E**dge) [16], which further strengthen the cooling in the edge region. The plasma cooling produces a contraction of plasma current profile and then leads to an increased current gradient and hence MHD destabilization inside the $q = 2$ surface. Eventually, a major disruption occurs [6, 17].

1.5.2 β limit

The maximum achievable plasma pressure (i.e. β limit) can be formulated as a limit on β_N which is the normalized β_t defined in Equation (1.7), given by

$$\beta_N \equiv \frac{\beta_t}{I_p/aB_0} \tag{1.13}$$

where I_p is the plasma current in MA, a is the minor radius in m and B_0 is the B_ϕ on axis. There are two β limits. The Sykes limit [18] of $\beta_N \leq 0.044$, or 4.4% is obtained from the analysis of largest β value stable to high n ballooning modes. The Troyon limit [19] of $\beta_N \leq 0.028$, or 2.8% comes from the more extensive study of MHD modes and plasma shapes. The Troyon limit is shown in Figure 1-5. Since Troyon limit is

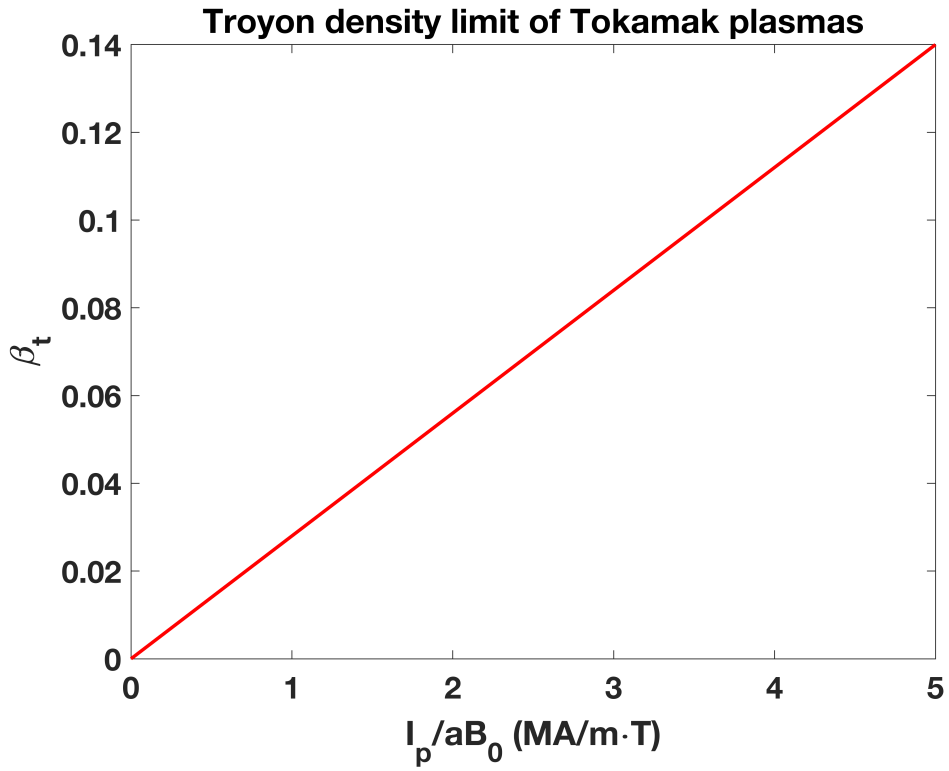

Figure 1-5: Troyon limit of Tokamak plasmas.

more limited, it is considered as the primary β limit which can also be written as

$$\beta_N \lesssim 4\ell_i \tag{1.14}$$

For highly elongated plasmas, like DIII-D plasma, the Troyon limit can occasionally be exceeded. A revised formula by Menard [20] shows Troyon limit is still valid for highly elongated plasmas if we redefine β_t using $\langle B_\phi^2 \rangle$ instead of B_0^2.

1.5.3 low-q disruption

The underlying physics mechanism of low-q disruption might be the increasing incompatibility of $m = 1$ modes and $m = 2$ modes stability as the safety factor near the edge, q_{95}, decreases. Since the safety factor at center q_0 is limited by $m = 1$ sawtooth instability, increasing I_p and hence decreasing q_{95} makes the q-profile more unstable. Without sawtooth instability, increasing plasma current will increase the current density on axis. However, the sawtooth crash limits q_0 by flattening the q-profile in the central region. Thus, increasing current eventually leads to the increased current gradient in the outer region of the plasma, in turn resulting in the growth of a tearing/kink instability. This sets a lower limit on the edge safety factor, $q_{95} \geq 2$. This

low-q limit on tokamak operation can possibly be exceeded by using active control of the corresponding MHD instabilities, as demonstrated on DIII-D plasma [21]. At present, most of existing tokamaks (and near future burning plasma tokamaks) tend to operate with $q_{95} \gtrsim 3$ to decrease the possibility of having disruptions.

1.5.4 Vertical displacement event (VDE)

Equation (1.10) shows that tokamak plasmas require an external vertical field to eliminate the horizontal outward expansion in the vessel. An active feedback control system is needed to sustain the quilibrium of the plasma. Similarly, as shown in [22], elongated plasmas are unstable to a gross vertical displacement. Due to this instability, tokamaks with elongated plasmas need an active feedback control system on the timescale that allows magnetic field to penetrate the vessel wall to hold the vertical position of the plasma [23]. If the feedback control system fails to correct a vertical perturbation, the perturbation grows and might lead to a *vertical displacement event* (VDE) in which the whole plasma moves up or down. VDEs usually set the limit on the maximum achievable elongation. The danger of VDEs became clear on JET where, in at least one case, forces of several hundred tons were generated on the vacuum vessel [24].

VDEs can occur before or after the thermal quench. If VDEs happen before the thermal quench, the hot plasma can make large contact with the wall of vessel which is called a *hot VDE*. These events lead to the cooling and recombining of the edge plasma immediately after it contact with the wall. The recombination process continues until the plasma shrinks such that $q_{95} \sim 2$ and then the plasma equilibrium is quickly destroyed. Once this happens, the toroidal field is trapped between the wall and the plasma which drives a large poloidal current flow in the conducting wall. At this stage, the outer flux surfaces intersect the wall over a "halo" region and the driven current is called the *halo currents*, as shown in Figure 1-6. The large electromagnetic force resulting from halo currents can be a threat to the components inside the vessel as well as to the vessel itself.

1.5.5 Tearing mode and Locked mode

In tokamak plasmas, although magnetic perturbations $\delta\mathbf{B}$ which bend field lines tend to be stabilized by magnetic tension force, the MHD equilibrium discussed in Section 1.4 can be unstable to some resonant perturbations. $\delta\mathbf{B}$ can be expressed using using Fourier components with $\delta\mathbf{B} \sim e^{i(m\theta+n\phi)}$. If a particular perturbation and a magnetic surface ψ satisfy

$$q(\psi) = \left|\frac{m}{n}\right| \tag{1.15}$$

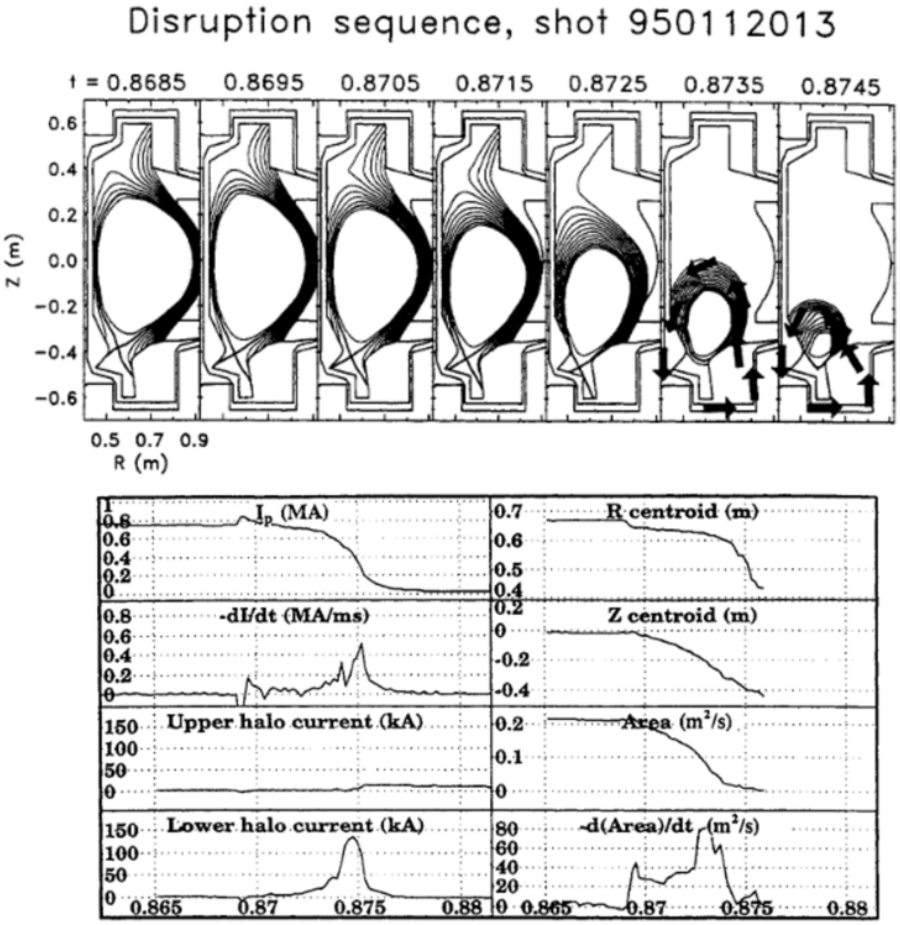

Figure 1-6: a VDE coinciding with current quench on Alcator C-Mod, from Bob Granetz [25]. Magnetic flux surface reconstructions with 1 ms time resolution are shown in the upper plot. The direction of driven halo currents are given by the arrows in the last two frames. In the lower plot, temporal evolution of several key plasma signals during the VDE are shown.

the wavefronts of the mode align along the field lines of the surface and it becomes resonant to the surface. In Equation (1.15), q is the safety factor (Equation (1.11)) and m and n are the poloidal and toroidal integer mode numbers, respectively, corresponding to the perturbation. Magnetic surfaces satisfying Equation (1.15) are called *rational surfaces* and the perturbations resonant on these surfaces are called (m, n) *modes*. The (m, n) mode has m poloidal wavelengths for every n toroidal wavelengths and thus higher (m, n) mode are more stable due to larger field line bending. The lowest (m, n) modes are the most dangerous resonant modes.

Among various MHD instabilities, the current driven $(2, 1)$ and $(3, 2)$ tearing modes [26] are known to be among the most relevant instabilities for tokamak disruptions. Usually, a slow change in MHD equilibrium pushes the current profile to marginal stability and finally across the tearing mode onset boundary. At the first stage, the tearing modes appear as low level oscillations. The phase velocity (i.e. frequency) of the mode can limit the mode growth as suggested by [15, 27, 28]. However, when the modes grow large enough, they can interact with external conductors like the vessel. The interaction between modes and vacuum vessel will add a torque to the plasma which decelerate and finally halt the mode propagation. When the frequency of the mode decreases to 0, the mode is locked to the vessel. This process of removal of the mode oscillation is known as the *mode locking* [29]. A example of mode locking on DIII-D is given by Figure 1-7.

Locked modes can sometimes be the result of an initially stationary 'born' locked mode due to the penetration of *error fields*. Error fields can induced by (a) the winding structure of coils, (b) the connections to the coils and (c) small misalignment of the coils [30]. Error fields can introduce perturbing deviations to the toroidal symmetry of the magnetic field which lead to the growth of $(2, 1)$ tearing modes. Usually error fields with an $m = 2$ radial component, B_r, of a few gauss can be enough to cause a locked mode, which presents a challenge for the design of future reactors.

1.5.6 Radiative Collapse and Impurities

Radiation loss of plasma energy can result in several different ways including atomic processes and the acceleration of charged particles. For the first one, the atomic line emission and recombination can lead to radiation. In steady state, considering the plasma with single impurity species, the atomic radiation given by impurities is

$$P_R = n_e n_I R_I(T_e) \tag{1.16}$$

where n_I is the impurity density and $R_I(T_e)$ depends on electron temperature as shown in [6]. The function $R_I(T_e)$ reaches maximum at $\sim$10 eV for lighter impurities

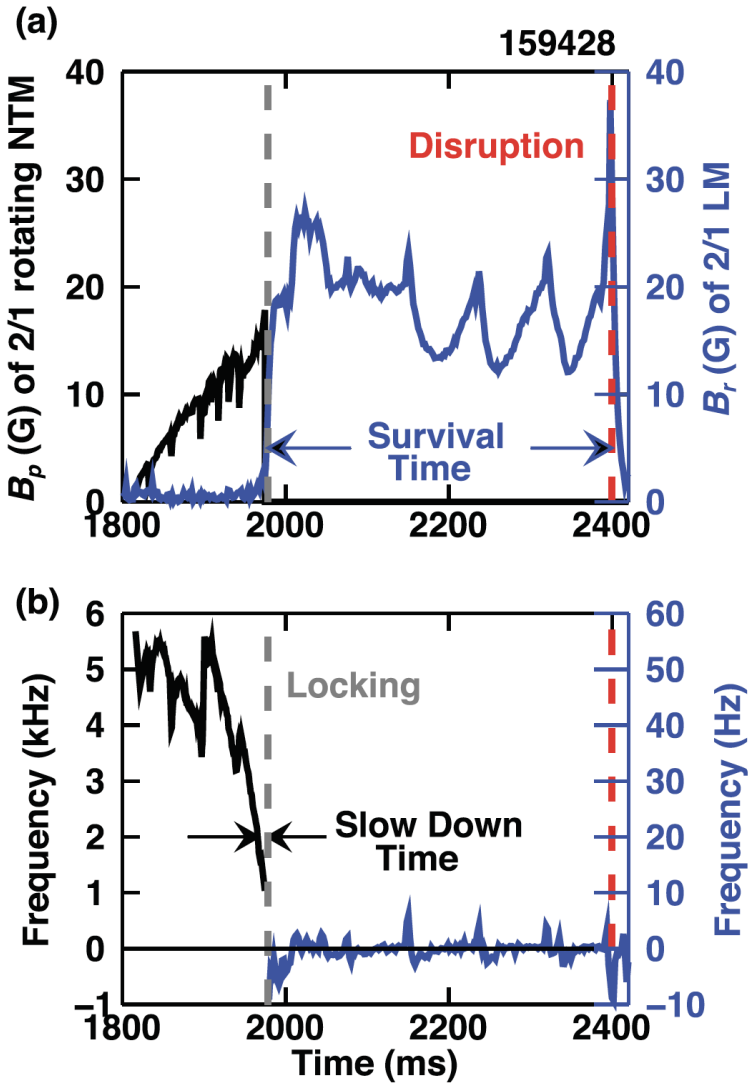

Figure 1-7: Locked mode lead to a disruption on DIII-D, from Ryan Sweeney [27]. The time traces of the amplitude of both the fast $(2, 1)$ neoclassical tearing mode (NTM, in black) and low frequency locked mode (LM, in blue) are shown in the upper plot. The fast NTM is locked at 1978.5 ms and the low frequency locked mode is detected at this point. In the lower plot, the frequency evolution of two modes are given. The slow down time (interval between mode rotating at 2 kHz) to lock) and the survival time (interval between mode locking and disruption) are marked in two plots.

and ∼1 keV for heavier impurities. Thus, high-Z impurities can radiate significantly at core region of the plasma and they must be stopped from entering the core. To this aim, most of modern tokamaks equipped with *divertor* can reduce the probability that impurities will penetrate through the scrape-off layer into the core plasma [31].

For the latter way of radiation loss, a mechanism called *Bremsstrahlung* radiation plays an important role. The Bremsstrahlung radiation is emitted by the the acceleration of a charged particle deflected by other charged particles. In plasmas, this can happen when electrons interact with ions through Coulomb collisions. For a plasma with multiple ion species, the Bremsstrahlung radiation power density is

$$P_B \propto Z_{eff} n_e^2 T_e^{1/2} \tag{1.17}$$

where $Z_{eff} = \frac{1}{n_e}\sum_j Z_j^2 n_j$ is the effective charge [4]. Clearly, the radiation becomes larger when the heavier impurity ions appear in the plasmas [32, 33].

When radiation power density from P_B and P_R is comparable to the input power densities, plasma energy balance can be broken, leading to a *radiative collapse*. Since the radiation power depends on impurity density, the influx of impurities from high-Z metal vessel wall (sometimes known as UFOs) can result in a rapid radiative collapse which often leads to quick disruptions. UFOs have been found to account for a large fraction of disruptions on high-Z metal wall tokamaks like Alcator C-Mod [34]. Impurity accumulation can also cause other radiation effects like the MARFE from Section 1.5.1.

1.6 Disruption Quench Phases and Consequences

1.6.1 Thermal and current quenches

After the break-up of magnetic surfaces, the final phases of disruptions start from the *thermal quench*, at which the plasma quickly loses all its thermal energy and deposits the energy onto the first wall of the vacuum vessel by conduction. Usually, a quick positive plasma current spike is observed at this point which is related to a rapid flattening of the radial plasma current profile. A typical current spike is shown in Figure 1-8. The sudden loss of thermal energy cools the entire plasma and significantly increases the resistivity $\eta \propto T_e^{-3/2}$ which then results in the the *current quench*. During the current quench, the plasma current decays to zero on the timescale $\tau_{CQ} \propto S/\eta$, where S is the plasma cross section area [35]. The typical timescale for both thermal and current quenches on tokamaks is of the order of few milliseconds [36].

Rapid energy loss due to the thermal and current quench can seriously damage the

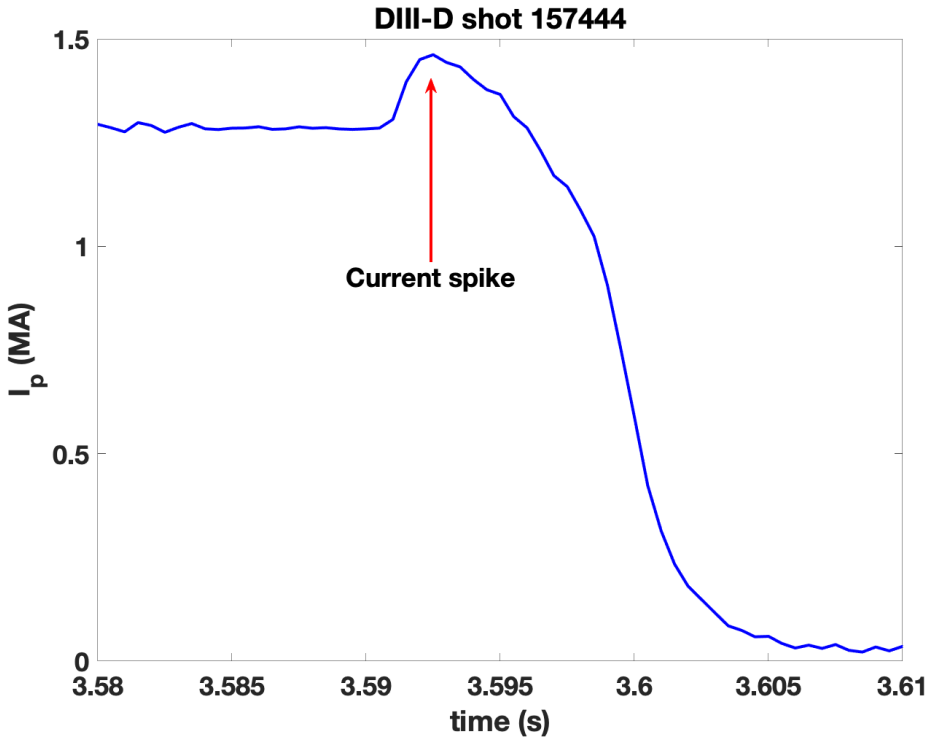

Figure 1-8: The current spike before current quench during a major disruption.

tokamak. Heat deposited to the first wall via thermal quench or hot VDEs becomes even more damaging when the heat load distribution is asymmetric [37]. Furthermore, the rapid change of plasma current during the current quench can drive currents like halo currents and eddy currents in the conducting components of the vessel wall. The induced currents then give rise to large Lorentz forces which also have the potential to damage the device. Again, toroidal asymmetries of the induced halo currents can increase the damage [38, 39].

1.6.2 Runaway Electrons

Another important consequence of disruptions is the potential for the formation of a large number of relativistic *runaway electrons* flowing toroidally. The formation of runaway electrons is discussed in [40]. In short, the Coulomb collisions between electrons traveling along the magnetic field with background electrons add a drag force to the traveling electrons. When the force due to the parallel component of electric field $E_{\|}$ exceeds the drag force, electrons in the high-energy tail of the velocity distribution can be continuously accelerated and reach relativistic energies. The generated so-called runaway electrons can transfer energy to background electrons via knock-on collisions which further gives rise to an avalanche process [41] and finally leading to

a "runaway beam" of relativistic electrons. The generated runaway beam sometimes carries a large portion of total plasma current. If the control of beam is lost, the runaway electron beam can directly hit the plasma facing components locally and generate a very large volumetric heat in the plasma facing components. This large heat can then melt the tiles of the first wall and damage plasma facing components [42–44].

1.7 Disruption Avoidance and Disruption Mitigation

As suggested by [45, 46], near future burning-plasma tokamaks like ITER cannot withstand more than a few unmitigated, high current, high stored energy disruptions because such disruptions can threaten the integrity of the facility. Ideally, disruption frequency can be minimized by avoiding scenarios which come near to operational limits (as discussed in Section 1.5). Additionally, optimized current and pressure profiles can minimize the chance of having MHD mode onset [47].

In addition to the passive avoidance strategies mentioned above, disruptions can be actively avoided using real-time control systems. If the precursor events in the early stage of a disruption can be detected, the pre-programmed control systems can then take proper actions corresponding to the detected events which might steer plasma away from the unstable regime. Previous disruption avoidance studies have included stabilizing neoclassical tearing modes via electron cyclotron current drive (ECCD) [48] and preventing tearing mode locking by external driven rotating resonant magnetic perturbations (RMPs) [49]. However, due to the complexity of disruptions and the lack of a fast plasma simulator, determining the early unstable events and taking proper actions in real time can be challenging.

Finally, if the thermal quench cannot be avoided, a disruption mitigation system (DMS) needs to be triggered to alleviate the consequence of heat and electromagnetic loads from thermal and current quenches. The idea of most mitigation stategies is to cool the plasma by radiation which uniformly deposits thermal energy on the first wall to prevent the concentration of heat loads on the divertor [50]. Given that the radiation power increases with the density and charge of the impurity ions (see Section 1.5.6), different quantities and types of injected impurities give different radiation timescales to address the mitigation needs. Examples include shattered pellet injection (SPI) [51] which inject impurity pellets that will ionize in the desired region of the plasma and massive gas injection (MGI) by injection of a large quantity of deuterium or impurity gas into the plasma [52, 53]. In addition, a previous study has demonstrate the passive prevention of runaway electron beam formation during disruptions via a non-axisymmetric in-vessel coil [54].

Table 1.1: Tokamak design parameters [9, 55]

Parameter	Units	Alcator C-Mod	DIII-D	EAST
Major radius, R	m	0.67	1.66	1.70
Minor radius, a	m	0.21	0.67	0.40
Elongation[a], κ		1.8	2.01	2.0
Maximum toroidal magnetic field, B_0	T	8.0	2.2	3.5
Maximum toroidal plasma current, I_p	MA	2.0	2.0	1.0
Maximum flattop duration[b], $\Delta t_{flattop}$	s	1	6	1000
Average electron density, n_e (typical)	10^{20} m^{-3}	~ 1	~ 0.4	~ 0.4
Core electron temperature, T_{e0} (typical)	keV	~ 2	~ 3	~ 2
Auxiliary power[c], $P_{aux,max}$	MW	6	27	28
First wall material,		molybdenum	carbon	hybrid[d]

[a]Elongation is defined at the plasma separatrix
[b]Flattop is the period of constant plasma current
[c]Maximum auxiliary heating power coupled to the plasma
[d]lower divertor: carbon, middle wall: molybdenum, upper divertor: tangsten

1.8 Alcator C-Mod, DIII-D, and EAST

Data driven analysis presented in this thesis is based on the data from three tokamaks, *Alcator C-Mod*, *DIII-D*, and *EAST*. Alcator C-Mod was a compact, high-magnetic field, diverted tokamak with a molybdenum first wall, located at the MIT Plasma Science and Fusion Center. DIII-D is a medium size, diverted tokamak with a carbon wall, located at General Atomics in San Diego. EAST is a medium size, superconducting, diverted tokamak with a hybrid first wall (lower divertor: carbon, middle wall: molybdenum, upper divertor: tangsten), located at Institute of Plasma Physics Chinese Academy of Sciences. C-Mod stopped operating in 2016 while DIII-D and EAST are still in operation. The main design parameters of these three tokamaks are summarized in Table 1.1 and there exists substantial difference in size and magnetic field between C-Mod and the other two tokamaks. In addition, as mentioned above, these three devices use different materials for the plasma facing components. The combination of these different characteristics covers a substantial fraction of ITER's features [45], although no existing device can, by itself, fully represent ITER at scale. In addition, since ITER will operate in a new regime that none of the current machine can approach, the new physics that will emerge in the ITER operational regime is not fully considered in present simulation codes and thus the extrapolation of current codes to ITER situation is uncertain. Different level of fidelity codes exist, but no whole plant modeling tool is a great challenge. A cross-machine study using data from these existing devices is nevertheless well-suited for an investigation of possible disruption prediction/avoidance solutions for ITER.

1.9 Outline

Magnetic confinement fusion energy via tokamaks is one of the most attractive schemes for future clean energy production. In order to produce net energy, tokamaks need to maintain a hot plasma cloud with high pressure and good confinement quantity as discussed in Section 1.2 which sometimes requires the tokamaks operate near several operational limits (see Section 1.5). Furthermore, the plasma's internal pressure and current profile and the magnetic topology can also give rise to certain plasma instabilities (see Section 1.5). Fast growing plasma instabilities usually lead to disruptions in which the plasma loses all its thermal and magnetic energy on the order of few milliseconds. The resulting consequences from disruptions can significantly threaten the integrity of next-generation reactors like ITER (see Section 1.6) and thus disruptions must be reliably predicted and avoided/mitigated on these next-step machines. This thesis proposes several data-driven methods for facilitating the prediction and avoidance of disruptions across multiple tokamaks based on the time traces of signals characterizing the state of the plasma. It also discusses the possible strategies to develop disruption prediction systems on future tokamaks.

In chapter 2, several methods proposed in previous studies for predicting and avoiding disruptions are reviewed with a focus on data-driven and machine learning methods. Then, in chapter 3, based on the conclusions from an unsupervised data exploration study, a hybrid deep-learning (HDL) model for general disruption prediction across multiple tokamaks is developed and the performance of the HDL model and other disruption prediction models are compared highlighting the advantages of the HDL model. In addition, several device-independent qualitative conclusions about cross machine disruption prediction are obtained via cross machine numerical disruption prediction studies with the HDL model using data from several tokamaks. To facilitate the development of disruption prediction on future machines like ITER, a scenario adaptive disruption prediction study via the HDL model that aims to accurately predict disruptions in high performance (HP) regimes using only low performance (LP) data from the target device is discussed in chapter 4. The ability to identify disruption precursors is important to inform the operator about the triggered disruption warnings from the predictor and is critical for disruption avoidance. To this end, an upgraded HDL model that incorporates predictive capability of various plasma unstable events including rotating modes, locked modes, H-to-L back transitions and radiative collapses is demonstrated in chapter 5. The upgraded HDL model gives longer warning times and better cross machine ability compared to the base HDL model and the same upgrading strategy can be applied to any neural network based disruption predictor. Finally, a symbolic boundary for predicting n=1 tearing mode (TM) onset across tokamaks, developed by data-driven methods, is discussed in

chapter 6. In chapter 7, all major results from this thesis and the main contributions from the author are summarized and the possible future research directions building on the results of this thesis are discussed.

References - Chapter 1

[1] IEA. World Energy Outlook 2020. Technical report, International Energy Agency, 2020.

[2] BP. Statistical Review of World Energy. Technical Report 69, British Petroleum, 2020.

[3] JGJ Olivier and JAHW Peters. Trends in global co2 and total greenhouse gas emissions. *PBL Netherlands Environmental Assessment Agency: The Hague, The Netherlands*, 2020.

[4] J. Freidberg. *Plasma Physics and Fusion Energy*. Cambridge University Press, 2007.

[5] J D Lawson. Some criteria for a power producing thermonuclear reactor. *Proceedings of the Physical Society. Section B*, 70(1):6–10, jan 1957.

[6] John Wesson. *Tokamaks; 4th ed.* International series of monographs on physics. Oxford Univ. Press, Oxford, 2011.

[7] VP Smirnov. Tokamak foundation in ussr/russia 1950–1990. *Nuclear fusion*, 50(1):014003, 2009.

[8] V Mukhovatov, M Shimada, A N Chudnovskiy, A E Costley, Y Gribov, G Federici, O Kardaun, A S Kukushkin, A Polevoi, V D Pustovitov, Y Shimomura, T Sugie, M Sugihara, and G Vayakis. Overview of physics basis for ITER. *Plasma Physics and Controlled Fusion*, 45(12A):A235–A252, Nov 2003.

[9] A. J. Creely and M. J. et al. Greenwald. Overview of the SPARC tokamak. *Journal of Plasma Physics*, 86(5):865860502, 2020.

[10] J. Freidberg. *Ideal MHD*. Cambridge University Press, 2014.

[11] Harold Grad and Hanan Rubin. Hydromagnetic equilibria and force-free fields. *Journal of Nuclear Energy (1954)*, 7(3-4):284–285, 1958.

[12] V. D. Shafranov. Plasma Equilibrium in a Magnetic Field. *Reviews of Plasma Physics*, 2:103, Jan 1966.

[13] M. Murakami, J.D. Callen, and L.A. Berry. Some observations on maximum densities in tokamak experiments. *Nuclear Fusion*, 16(2):347–348, apr 1976.

[14] Martin Greenwald, J. L. Terry, S. M. Wolfe, S. Ejima, M.G. Bell, S. M. Kaye, and G. H. Neilson. A new look at density limits in tokamaks. *Nuclear Fusion*, 28(12):2199–2207, 1988.

[15] P. C. De Vries, M. F. Johnson, B. Alper, P. Buratti, T. C. Hender, H. R. Koslowski, and V. Riccardo. Survey of disruption causes at JET. *Nuclear Fusion*, 51(5):053018, 2011.

[16] V. P. Lipschultz, R. La Bombard, P. Marmar, R. Pickrell, R. R. Terry, R. R. Watterson, and S. M. Wolfe. Marfe: An edge plasma phenomenon. *Nuclear Fusion*, 24(8):977–988, 1984.

[17] Q. Teng, D.P. Brennan, L. Delgado-Aparicio, D.A. Gates, J. Swerdlow, and R.B. White. A predictive model for the tokamak density limit. *Nuclear Fusion*, 56(10):106001, Jul 2016.

[18] A. Sykes, M. F. Turner, and S. Patel. Proceedings of the 11th european conference on controlled fusion and plasma physics. volume 2, page 363. European Physical Society.

[19] F. Troyon and R. Gruber. A semi-empirical scaling law for the β-limit in tokamaks. *Physics Letters A*, 110(1):29–34, 1985.

[20] J. E. Menard, M. G. Bell, R. E. Bell, D. A. Gates, S. M. Kaye, B. P. LeBlanc, R. Maingi, S. A. Sabbagh, V. Soukhanovskii, and D. Stutman. Aspect ratio scaling of ideal no-wall stability limits in high bootstrap fraction tokamak plasmas. *Physics of Plasmas*, 11(2):639–646, 2004.

[21] P. Piovesan, J. M. Hanson, P. Martin, G. A. Navratil, F. Turco, J. Bialek, N. M. Ferraro, R. J. La Haye, M. J. Lanctot, M. Okabayashi, C. Paz-Soldan, E. J. Strait, A. D. Turnbull, P. Zanca, M. Baruzzo, T. Bolzonella, A. W. Hyatt, G. L. Jackson, L. Marrelli, L. Piron, and D. Shiraki. Tokamak operation with safety factor $q_{95} < 2$ via control of MHD stability. *Phys. Rev. Lett.*, 113:045003, Jul 2014.

[22] M. Okabayashi and G. Sheffield. Vertical stability of elongated tokamaks. *Nuclear Fusion*, 14(2):263–265, apr 1974.

[23] E.A. Lazarus, J.B. Lister, and G.H. Neilson. Control of the vertical instability in tokamaks. *Nuclear Fusion*, 30(1):111–141, jan 1990.

[24] V Riccardo, P Noll, and SP Walker. Forces between plasma, vessel and tf coils during avdes at jet. *Nuclear fusion*, 40(10):1805, 2000.

[25] R.S Granetz, I.H Hutchinson, J Sorci, J.H Irby, B LaBombard, and D Gwinn. Disruptions and halo currents in Alcator C-Mod. *Nuclear Fusion*, 36(5):545–556, May 1996.

[26] P.H. Rutherford. Tearing modes in tokamaks. In B. Coppi, G.G. Leotta, D. Pfirsch, R. Pozzoli, and E. Sindoni, editors, *Physics of Plasmas Close to Thermonuclear Conditions*, pages 129–142. Pergamon, 1981.

[27] R. Sweeney, W. Choi, R.J. La Haye, S. Mao, K.E.J. Olofsson, and F.A. Volpe. Statistical analysis of m / n = 2/1 locked and quasi-stationary modes with rotating precursors at DIII-D. *Nuclear Fusion*, 57(1):016019, Jan 2017.

[28] S.P. Gerhardt, D.S. Darrow, R.E. Bell, B.P. LeBlanc, J.E. Menard, D. Mueller, A.L. Roquemore, S.A. Sabbagh, and H. Yuh. Detection of disruptions in the high- β spherical torus NSTX. *Nuclear Fusion*, 53(6):063021, 6 2013.

[29] M.F.F. Nave and J.A. Wesson. Mode locking in tokamaks. *Nuclear Fusion*, 30(12):2575–2583, dec 1990.

[30] R Fitzpatrick. Interaction of tearing modes with external structures in cylindrical geometry (plasma). *Nuclear Fusion*, 33(7):1049–1084, jul 1993.

[31] Alberto Loarte. Effects of divertor geometry on tokamak plasmas. *Plasma Physics and Controlled Fusion*, 43(6):R183–R224, may 2001.

[32] S K Rathgeber, R Fischer, S Fietz, J Hobirk, A Kallenbach, H Meister, T Pütterich, F Ryter, G Tardini, and E Wolfrum and. Estimation of profiles of the effective ion charge at ASDEX upgrade with integrated data analysis. *Plasma Physics and Controlled Fusion*, 52(9):095008, aug 2010.

[33] M B Chowdhuri, R Manchanda, J Ghosh, K A Jadeja, Kaushal M Patel, Vinay Kumar, Ketan M Patel, P K Atrey, Y Shankara Joisa, S B Bhatt, and R L Tanna and. Investigation of the behavior of effective charge of ADITYA tokamak plasmas. *Plasma Physics and Controlled Fusion*, 62(3):035015, Feb 2020.

[34] J. Zhu, C. Rea, K. J. Montes, R. Granetz, R. Sweeney, and R. A. Tinguely. Hy-

brid deep learning architecture for general disruption prediction across tokamaks. *Nuclear Fusion*, 61(2):026007, 2020.

[35] *Fusion Physics*. Non-serial Publications. INTERNATIONAL ATOMIC ENERGY AGENCY, Vienna, 2012.

[36] V. Riccardo and A. Loarte. Timescale and magnitude of plasma thermal energy loss before and during disruptions in JET. *Nuclear Fusion*, 45(11):1427–1438, 2005.

[37] Jochen Linke, Juan Du, Thorsten Loewenhoff, Gerald Pintsuk, Benjamin Spilker, Isabel Steudel, and Marius Wirtz. Challenges for plasma-facing components in nuclear fusion. *Matter and Radiation at Extremes*, 4(5):056201, 2019.

[38] S.N. Gerasimov, P. Abreu, M. Baruzzo, V. Drozdov, A. Dvornova, J. Havlicek, T.C. Hender, O. Hronova, U. Kruezi, X. Li, T. Markovič, R. Pánek, G. Rubinacci, M. Tsalas, S. Ventre, F. Villone, and L.E. Zakharov and. JET and COMPASS asymmetrical disruptions. *Nuclear Fusion*, 55(11):113006, sep 2015.

[39] C. E. Myers, N. W. Eidietis, S. N. Gerasimov, S. P. Gerhardt, R. S. Granetz, T. C. Hender, and G. Pautasso. A multi-machine scaling of halo current rotation. *Nuclear Fusion*, 58(1), 2018.

[40] Boris N. Breizman, Pavel Aleynikov, Eric M. Hollmann, and Michael Lehnen. Physics of runaway electrons in tokamaks. *Nuclear Fusion*, 59(8):083001, jun 2019.

[41] M.N Rosenbluth and S.V Putvinski. Theory for avalanche of runaway electrons in tokamaks. *Nuclear Fusion*, 37(10):1355–1362, Oct 1997.

[42] Allen H. Boozer. Theory of tokamak disruptions. *Physics of Plasmas*, 19(5), 2012.

[43] B. Bazylev, G. Arnoux, W. Fundamenski, Yu. Igitkhanov, and M. Lehnen. Modeling of runaway electron beams for JET and ITER. *Journal of Nuclear Materials*, 415(1, Supplement):S841–S844, 2011. Proceedings of the 19th International Conference on Plasma-Surface Interactions in Controlled Fusion.

[44] C. Reux, V. Plyusnin, B. Alper, D. Alves, B. Bazylev, E. Belonohy, A. Boboc, S. Brezinsek, I. Coffey, J. Decker, P. Drewelow, S. Devaux, P.C. de Vries, A. Fil, S. Gerasimov, L. Giacomelli, S. Jachmich, E.M. Khilkevitch, V. Kiptily, R. Koslowski, U. Kruezi, M. Lehnen, I. Lupelli, P.J. Lomas, A. Manzanares, A. Martin De Aguilera, G.F. Matthews, J. Mlynář, E. Nardon, E. Nilsson, C. Perez von Thun, V. Riccardo, F. Saint-Laurent, A.E. Shevelev, G. Sips, and C. Sozzi and. Runaway electron beam generation and mitigation during disruptions at JET-ILW. *Nuclear Fusion*, 55(9):093013, Aug 2015.

[45] T.C. Hender, J.C. Wesley, J. Bialek, A. Bondeson, A.H. Boozer, R.J. Buttery, A. Garofalo, T.P. Goodman, R.S. Granetz, Y. Gribov, O. Gruber, M. Gryaznevich, G. Giruzzi, S. Günter, N. Hayashi, P. Helander, C.C. Hegna, D.F. Howell, D.A. Humphreys, G.T.A. Huysmans, A.W. Hyatt, A. Isayama, S.C. Jardin, Y. Kawano, A. Kellman, C. Kessel, H.R. Koslowski, R.J. la Haye, E. Lazzaro, Y.Q. Liu, V. Lukash, J. Manickam, S. Medvedev, V. Mertens, S.V. Mirnov, Y. Nakamura, G. Navratil, M. Okabayashi, T. Ozeki, R. Paccagnella, G. Pautasso, F. Porcelli, V.D. Pustovitov, V. Riccardo, M. Sato, O. Sauter, M.J. Schaffer, M. Shimada, P. Sonato, E. J. Strait, M. Sugihara, M. Takechi, A.D.

Turnbull, E. Westerhof, D.G. Whyte, R. Yoshino, and H. Zohm. Chapter 3: Mhd stability, operational limits and disruptions. *Nuclear Fusion*, 47(6):S128–S202, June 2007.

[46] M. Lehnen. Plasma disruption management in ITER. In *26th IAEA Fusion Energy Conference IAEA*, pages EX/P6–39, 2016.

[47] F. Turco, T.C. Luce, W. Solomon, G. Jackson, G.A. Navratil, and J.M. Hanson. The causes of the disruptive tearing instabilities of the ITER baseline scenario in DIII-D. *Nuclear Fusion*, 58(10):106043, Sep 2018.

[48] H Zohm, G Gantenbein, G Giruzzi, S Günter, F Leuterer, M Maraschek, J Meskat, AG Peeters, W Suttrop, D Wagner, et al. Experiments on neoclassical tearing mode stabilization by eccd in asdex upgrade. *Nuclear Fusion*, 39(5):577, 1999.

[49] D. Li, Q. Yu, Y. Ding, N. Wang, F. Hu, R. Jia, L. Peng, B. Rao, Q. Hu, H. Jin, M. Li, L. Zhu, Z. Huang, Z. Song, S. Zhou, J. Li, Y. He, Q. Zhang, W. Zhang, J. Dong, D. Han, W. Zheng, A. A. Bala, K. Yu, and Y. Liang and. Disruption prevention using rotating resonant magnetic perturbation on J-TEXT. *Nuclear Fusion*, 60(5):056022, Apr 2020.

[50] M. Lehnen, K. Aleynikova, P.B. Aleynikov, D.J. Campbell, P. Drewelow, N.W. Eidietis, Yu. Gasparyan, R.S. Granetz, Y. Gribov, N. Hartmann, E.M. Hollmann, V.A. Izzo, S. Jachmich, S.-H. Kim, M. Kočan, H.R. Koslowski, D. Kovalenko, U. Kruezi, A. Loarte, S. Maruyama, G.F. Matthews, P.B. Parks, G. Pautasso, R.A. Pitts, C. Reux, V. Riccardo, R. Roccella, J.A. Snipes, A.J. Thornton, and P.C. de Vries. Disruptions in ITER and strategies for their control and mitigation. *Journal of Nuclear Materials*, 463:39–48, 2015. PLASMA-SURFACE INTERACTIONS 21.

[51] D. Shiraki, N. Commaux, L. R. Baylor, N. W. Eidietis, E. M. Hollmann, C. J. Lasnier, and R. A. Moyer. Thermal quench mitigation and current quench control by injection of mixed species shattered pellets in DIII-D. *Physics of Plasmas*, 23(6):062516, 2016.

[52] T. C. Jernigan, L. A. Baylor, S. K. Combs, D. A. Humphreys, P. B. Parks, and J. C. Wesley. Massive gas injection systems for disruption mitigation on the DIII-D tokamak. In *21st IEEE/NPS Symposium on Fusion Engineering SOFE 05*, pages 1–3, 2005.

[53] M Lehnen, A Alonso, G Arnoux, N Baumgarten, SA Bozhenkov, S Brezinsek, M Brix, T Eich, SN Gerasimov, A Huber, et al. Disruption mitigation by massive gas injection in jet. *Nuclear fusion*, 51(12):123010, 2011.

[54] RA Tinguely, VA Izzo, DT Garnier, A Sundström, K Särkimäki, O Embréus, T Fülöp, RS Granetz, M Hoppe, I Pusztai, et al. Modeling the complete prevention of disruption-generated runaway electron beam formation with a passive 3d coil in sparc. *Nuclear Fusion*, 61(12):124003, 2021.

[55] M. Greenwald, A. Bader, S. Baek, M. Bakhtiari, H. Barnard, W. Beck, W. Bergerson, I. Bespamyatnov, P. Bonoli, D. Brower, D. Brunner, W. Burke, J. Candy, M. Churchill, I. Cziegler, A. Diallo, A. Dominguez, B. Duval, E. Edlund, P. Ennever, D. Ernst, I. Faust, C. Fiore, T. Fredian, O. Garcia, C. Gao, J. Goetz, T. Golfinopoulos, R. Granetz, O. Grulke, Z. Hartwig, S. Horne, N. Howard,

A. Hubbard, J. Hughes, I. Hutchinson, J. Irby, V. Izzo, C. Kessel, B. LaBombard, C. Lau, C. Li, Y. Lin, B. Lipschultz, A. Loarte, E. Marmar, A. Mazurenko, G. McCracken, R. McDermott, O. Meneghini, D. Mikkelsen, D. Mossessian, R. Mumgaard, J. Myra, E. Nelson-Melby, R. Ochoukov, G. Olynyk, R. Parker, S. Pitcher, Y. Podpaly, M. Porkolab, M. Reinke, J. Rice, W. Rowan, A. Schmidt, S. Scott, S. Shiraiwa, J. Sierchio, N. Smick, J. A. Snipes, P. Snyder, B. Sorbom, J. Stillerman, C. Sung, Y. Takase, V. Tang, J. Terry, D. Terry, C. Theiler, A. Tronchin-James, N. Tsujii, R. Vieira, J. Walk, G. Wallace, A. White, D. Whyte, J. Wilson, S. Wolfe, G. Wright, J. Wright, S. Wukitch, and S. Zweben. 20 years of research on the Alcator C-Mod tokamak. *Physics of Plasmas*, 21(11):110501, 2014.

Chapter 2

Disruption Prediction and Disruption Avoidance Models

To implement the avoidance and mitigation strategies discussed in Section 1.7, disruption prediction models with high accuracy need to be developed. To simply trigger the DMS, a typical disruption prediction model only has to generate a threshold-based signal that determines whether or not a mitigation system (like SPI or MGI) should be fired. To this aim, a successful model needs to achieve both high true positive rate (TPR, the successfully detected disruptions as a fraction of all disruptions) and high true negative rate (TNR, the successfully predicted non-disruptive shots as a fraction of all non-disruptive shots). the model should also give sufficiently long warning time to enable a successful mitigation (at least a few tens of milliseconds). For the ITER DMS, the DMS trigger requires a warning time greater than $\sim$30 ms with TPR as high as $\sim 99\%$ and TNR as high as $\sim 98\%$ during high performance operation [1].

As for the required avoidance system model, it not only needs to forecast the beginning of instability onset, but also needs to characterize the underlying precursor events that lead to the unstable phase and automatically choose the proper actions that can push the plasma back to a safer operational regime. Moreover, the algorithm should give a long enough warning time for chosen actions to be effective such that the upcoming disruption can be avoided (usually more than a few hundreds of milliseconds).

Disruption prediction models are required to secure the success of near future burning-plasma tokamaks, including ITER [2]. To date, disruption prediction has been studied through two main approaches: data-driven versus physics-driven (model-based). In this chapter, these two major branches of disruption prediction models are both reviewed in detail (see Section 2.1 and Section 2.2) and the performance comparison between them is discussed. The challenge of cross-machine disruption prediction for data-driven models is then discussed in Section 2.3, motivating the

studies presented in the following chapters of this thesis.

2.1 Physics-driven Approaches

Developing an integrated physical model based on first principles for disruption avoidance has been shown to be very difficult [1]. For example, a comprehensive physics-based model for MHD instabilities needs to solve the plasma equilibrium equation with magnetic islands based on measured magnetic signals as boundary conditions, and predict the island growth rate with small uncertainties. In addition, all these calculations must be done quickly enough such that the results can be used for real-time deployment [3]. Although significant efforts have been made along this direction [4], a code meeting all such requirements is still unavailable. Moreover, physics understanding about some transport phenomena is still inadequate. Therefore, most current model-based approaches for disruption prediction only consider a subset of disruption related unstable events, as for example described in Section 1.5.

The simplest model-based approach is the threshold-based method. For this approach, an alarm is triggered if one or more measured plasma signals exceed certain thresholds. For example, the measured values of relevant parameters like the plasma current, the electron density, and $(2, 1)$ mode amplitude can be used to initiate a safe shutdown action [5]. Locked mode amplitude measurements have been used to trigger MGI on both ASDEX Upgrade [6] and JET [7]. By fitting a scaling law for the locked mode amplitude prior to the thermal quench time on several existing tokamaks, this locked mode threshold model can possibly be extrapolated to future devices [1]. Furthermore, a statistical study about mode locking on DIII-D has shown the relation between ℓ_i/q_{95} and disruptive locked modes [8].

Modular approaches that combine identifiers of several disruption related unstable events have also been studied. For example, a multi-layered disruption predictor that combines several independent simple models for multiple unstable events has been implemented on ASDEX Upgrade [9]. Recently, robust off-normal event handling algorithms have been developed on DIII-D [10], TCV [11], and ASDEX Upgrade [12]. These algorithms incorporate multiple event identification modules and send outputs from event detectors to integrated control and actuator management systems to ameliorate the detected instabilities.

Models based on path-oriented analysis have also been developed. A recent review is presented in [4]. These algorithms focus on several common unstable event chains, as presented in Section 1.5, and output plasma proximity to a disruptive boundary as a function of time. The state of the art of these path-oriented analysis models is the Disruption Event Characterization and Forecasting (DECAF) algorithm [13], that combines several physics-driven models, each detecting different unstable events, and

generates an overall instability level. The modules incorporated in DECAF include a resistive wall mode detector [14], a rotating MHD mode detector [15], and an ELM detector [16]. DECAF has been implemented in real-time on KSTAR [17]. Other examples of path-oriented analysis models include a density limit disruption avoidance algorithm [18] and a model that focuses on disruptions that are caused by impurity influx and NTMs [19]. In general, the model-based approaches can provide early warnings of well studied common unstable events that can lead to disruptions on existing tokamaks. However, due to the limited knowledge of some unstable events, and the possible new physics like strong alpha heating on next-generation devices, the extrapolation of physics-driven approaches to future devices has significant uncertainty. In addition, the adoption of a physics-driven model is limited by our understanding of the underlying physics, and so far has not been done on a fast timescale (inter-discharge adaption).

2.2 Data-driven Approaches

The basic logic of disruption prediction and avoidance models is to construct a function $f(\mathbf{x})$ that maps an input vector $\mathbf{x}$ representing the current (and/or historical) plasma state, to an output vector $\mathbf{y}$ quantifying the risk level of different unstable events and/or disruption. For the physics-driven approaches, the mapping function $f(\mathbf{x})$ is explicitly designed purely based on the knowledge of the relevant plasma physics equations.

In the case of data-driven approaches, the function $f(\mathbf{x})$ is chosen from a large class of functions during a *training* process based on the functions' statistical performance on a large amount of historical data. The function class, also called hypothesis space, is usually sufficiently large to fit any continuous function closely enough and it is not limited by the set of functions that relate to physics models. The fitted function can then be applied to **unseen** or **new** data through a *testing* process to evaluate its generalization ability. This is the key difference between machine learning and a traditional optimization approach. For data-driven approaches, finding a proper model becomes an optimization and generalization problem rather than a physics problem. Nowadays, data-driven methods are also known as *machine learning* under the more general ensemble of *artificial intelligence* techniques.

The term machine learning was proposed by Arthur Samuel from IBM in 1959. Since then, this field has greatly advanced with a large variety of machine learning models and theoretical studies about training, understanding and deploying the models. Moreover, the roughly one trillion-fold increase in computational power since 1956 further supports the expansion of machine learning studies. Today, machine learning models are used in nearly every field in science and engineering. Fusion research has

also taken advantage of progress in machine learning over the past two decades with lmany fusion problems being studied via machine learning methods [20]. In this section, two major branches of machine learning, *supervised* and *unsupervised* learning algorithms are introduced. A representative of machine learning models, the *artificial neural network* or simply neural network, is then discussed in detail.

2.2.1 Supervised learning

In the case of supervised learning, the task is to learn a function that maps an input $\mathbf{x}$ to a target output $\mathbf{y}$ given example input-output pairs. Depending on the types of output, the task is called a *classification* problem if the output space is discrete (usually finite) or a *regression* problem if the output space is continuous (like the real line $\mathbb{R}$). A regression task is typically mapped as follows:

$$\mathbf{y} = f_{true}(\mathbf{x}) + \epsilon \tag{2.1}$$

where f_{true} is the ground truth function mapping the input vector $\mathbf{x}$ to the output vector $\mathbf{y}$ and ϵ is an independent random vector describing the noise. The classification tasks can also be formalized as probabilistic problems trying to find the conditional probability $\mathbb{P}(\mathbf{y}|\mathbf{x})$. If this $\mathbb{P}(\mathbf{y}|\mathbf{x})$ is determined, then the output $\mathbf{y} = f(\mathbf{x})$ is simply the $\mathbf{y}$ that has the largest probability given $\mathbf{x}$, i.e. the $\mathbf{y}$ that maximizes $\mathbb{P}(\mathbf{y}|\mathbf{x})$ ($\arg\max_{\mathbf{y}} \mathbb{P}(\mathbf{y}|\mathbf{x})$). Supervised learning algorithms essentially seek a function from the hypothesis space that can best fit the ground truth f_{true} based on the training examples.

A typical example of supervised machine learning tasks is the handwritten digit classification (see Figure 2-1) which is discussed in [21]. In the MNIST dataset, the dataset of handwritten digits, the input $\mathbf{x}$ is a 784×1 vector representing the 28×28 pixel image of handwritten digit, and the output $\mathbf{y}$ is a possible integer from digits $0 - 9$. The function $f(\mathbf{x})$ is chosen by maximizing the classification accuracy over N training examples $X = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$. In addition, the raw handwritten data (pictures) was *pre-processed* such that the digit is centered in the frame for each of the images in the dataset (see Figure 2-1) and each digit has roughly the same width and height with the same pixel intensity range $(0 - 255)$. The pre-processing of the dataset greatly reduces the difficulty of finding the function $f(\mathbf{x})$. The pre-processing of the raw data to extract more differentiable variables (features) is known as *feature engineering* [22].

Returning to the disruption prediction problem, different raw plasma signals can range over many orders of magnitude. Therefore, the raw plasma signals have to be pre-processed properly before being used in the machine learning algorithms. More-

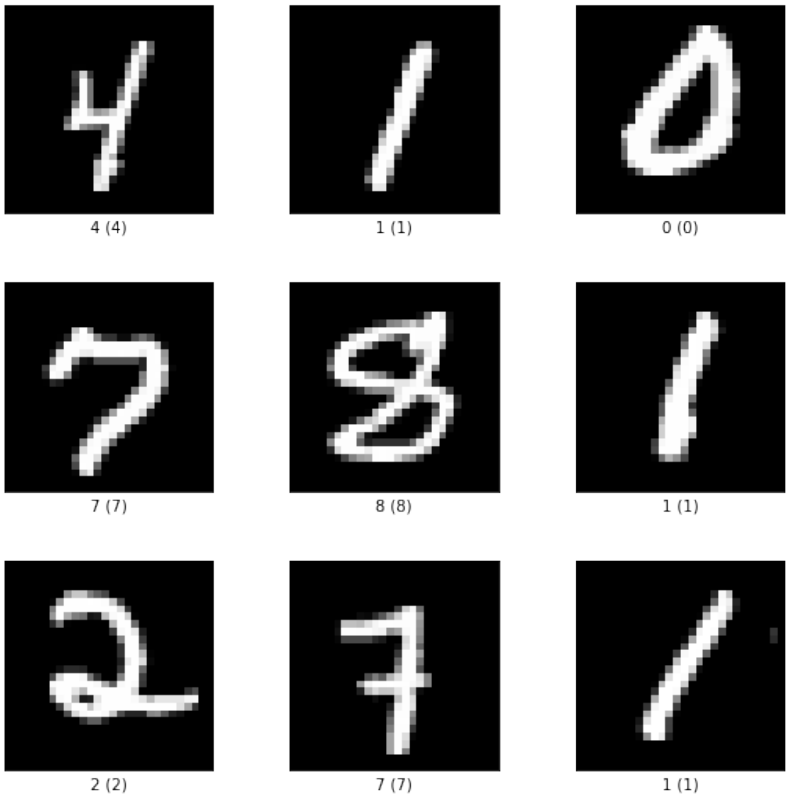

Figure 2-1: Examples of handwritten digits from the MNIST [23] dataset

over, several plasma signals from multiple diagnostics can be combined as a combo based on the subject matter expert's knowledge of disruption physics to strengthen the efficacy of input features. Finally, the target output $\mathbf{y}$ of the disruption prediction task is not uniquely defined. Different models can be trained to predict the time until disruption of the current plasma state (regression) or to classify the current plasma state is either stable or unstable (classification).

As mentioned before, the key difference between machine learning and optimization can be reduced to consideration of the generalization capabilities. Machine learning aims at minimizing the loss on unseen test data, while optimization is concerned with maximizing performance on the training set. Considering a fitted model $f(\mathbf{x})$, the mean square error (MSE) of the fitted model on a single test sample $\mathbf{x}_0$ can be defined as

$$Error(\mathbf{x}_0) = \mathbb{E}[(\mathbf{y} - f(\mathbf{x}_0))^2] \tag{2.2}$$

Substituting Equation (2.1) into Equation (2.2), with the assumptions that ϵ is independent of input feature $\mathbf{x}_0$, and has mean of zero and standard deviation σ_ϵ, the error can be rewritten as [24]:

$$Error(\mathbf{x}_0) = \Big(\mathbb{E}[f(\mathbf{x}_0)] - f_{true}(\mathbf{x}_0)\Big)^2 + \mathbb{E}\Big[\Big(f(\mathbf{x}_0) - \mathbb{E}[f(\mathbf{x}_0]\Big)^2\Big] + \sigma_\epsilon^2 \tag{2.3}$$

Equation (2.3) is called the bias-variance decomposition of the mean squared error. The first term in is the squared difference between the true value and estimated mean (squared *bias*) and the bias on the training set always decreases with increasing model complexity, because a more complex model can better fit the training data. The second term is the *variance*, which usually grows with increasing model complexity [1]. Ideally, machine learning algorithms should find a model that simultaneously minimizes both bias and variance. However, these two goals usually conflict. High variance models might fit the training set well but have higher chance to *overfit* to noisy or unrepresentative training data. In contrast, high bias models are usually simpler, but are at risk of missing important structures in the training data (i.e. *underfit*). A schematic plot that shows underfitting vs. optimal fitting vs. overfitting is given in Figure 2-2. In addition, there is an irreducible error of any fitted model coming from the third term in Equation (2.3), which is due to the measurement errors and randomness in the input/output data. In practice, another *validation set* is used (besides the training and testing sets) to reduce the bias and variance through tuning the model performance on this third dataset. The final tuned model is then applied to the testing set to evaluate its performance. An example of a typical general workflow

[1] models with high variance are 'complex' to some extent but a complex model does not necessarily have large variance [25]

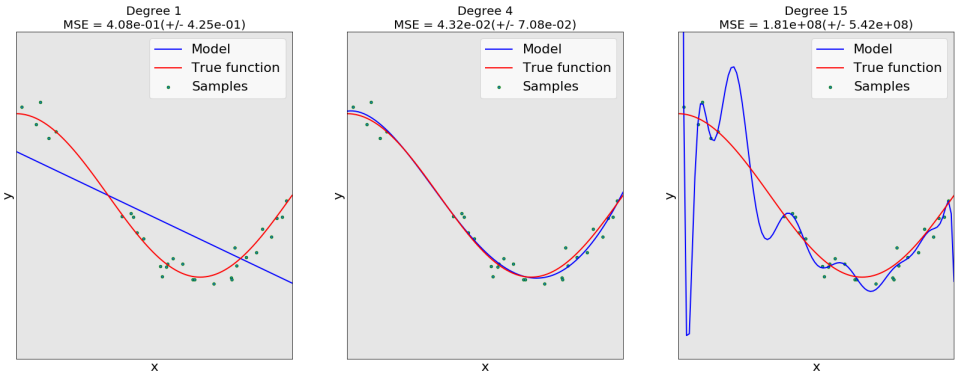

Figure 2-2: This schematic plot demonstrates the problems of underfitting and overfitting. From left to right, we show the cases of underfit, good fit and normal fit. The data points, the underlying true function (a cosine function) and the fitted model are given in each plot. It is clear that a linear function is not sufficient to fit the data points. This is called underfit. A polynomial of degree 4 approximates the true function almost perfectly. However, for higher degrees the model leads to an overfit.

diagram for machine learning is shown in Figure 2-3.

As mentioned before, disruption predictors for next-generation tokamaks like ITER must achieve very high accuracy to secure the success of the project. So far, various data-driven disruption prediction algorithms based on supervised machine learning have been developed on multiple tokamaks and shown attractive prediction performance. Examples of these algorithms include ensemble learning algorithm [27–29], support vector machines (SVM) [30–32], logistic regression (LR) [33] and neural networks (NN) [34, 35]. Several recent studies have incorporated both a physics-driven paradigm Section 2.1 and a data-driven idea to the same model to achieve better predictions, as well as improved interpretability [36]. Furthermore, revolu-

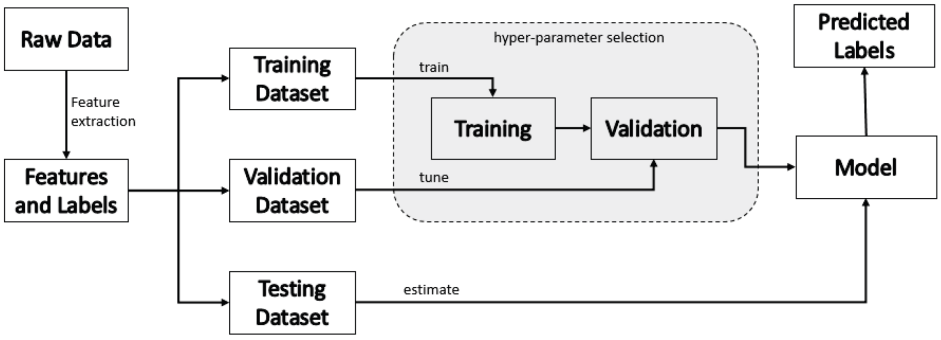

Figure 2-3: Example of a general machine learning workflow diagram from [26]

tionary advances in *deep learning* and significant expansion in computational power motivated recent deep neural network based disruption prediction studies [37–40]. These deep learning algorithms not only achieve state-of-the-art prediction performance but also are found to show potential for acquiring a general representation of experimental data that can be used in cross-machine applications. Although significant progress has been made in this direction, no data-driven approach has yet to achieve the prediction accuracy that is required by ITER, and the cross machine adoption of developed predictors is still a great challenge for both physics-driven and data-driven methods.

2.2.2 Unsupervised learning

For unsupervised learning, no labels $\mathbf{y}$ are given to the learning algorithm. The goal of unsupervised learning is to find hidden patterns or groupings in the data without the need for human labelling. Its ability to identify hidden data structures makes it an ideal method for exploratory analysis of high-dimensional plasma signals; the alternative, manual labeling by a fusion expert, is highly labor intensive, and only limited amounts of data can be practially processed. The unsupervised learning algorithm is mainly used for three tasks: clustering, association and dimensionality reduction. Examples of unsupervised learning algorithms previously applied to disruption studies include: generative topographical maps (GTM) [41]; k-means clustering [42]; t-SNE based dimensionality reduction [38]; self-organizing maps [43]; and variational autoencoders (VAE) based dimensionality reduction [44]. Some of these models have achieved similar accuracy as compared to the supervised learning algorithm discussed in Section 2.2.1.

Besides being directly used for disruption prediction, the plasma representations gleaned from an unsupervised learning algorithm can also be used to facilitate the design of more accurate supervised learning models. For example, in [38], the t-SNE clustering of plasma signals has shown the advantage of sequence-based plasma representation which further motivates the design of a supervised deep learning model. In addition, combining the learned 'latent space', with labelled unstable event information, can provide models for disruption avoidance. For example, the GTM, developed in [41], has been found to generate a boundary in a transformed 2D latent space that separates core and edge radiative collapses. The proximity of the plasma to the identified boundary in the low-dimensional latent space can be used to analyze the evolution of the risks of growth of certain instabilities over time, and to inform the plasma control system.

2.2.3 Artificial Neural Network

An artificial neural network, (or more simply referred to as a neural network), is a machine learning model that is inspired by, and somewhat analagous to the biological network of neurons in the animal brain. Neural network models are comprised of interconnected layers, with each layer consisting of nodes, called artificial neurons, that mimic neurons in the brain. Most of today's neural networks use "feed-forward" models, which means the information moves in only one direction through the networks. These feed-forward neural networks typically have one input layer from which the data is sent into the model, one output layer that generates the final output of the model, and several intermediate "hidden" layers. Each neuron in the network is connected to several neurons in the preceding layer, from which it receives data (except for neurons in the input layer) and in turn is connected to several neurons in the following layer to which it sends data (see Figure 2-4 for the visualization of this structure). Each neuron has a number called "bias" and each connection between two neurons has a number called "weight". In the operation of the network, a neuron receives the biases from each of its respective preceding neurons, multiplies those by the respective weights associated with each connection, and summed up by the neuron, yielding a scalar. This is added with bias of the neuron and then passed through a predefined nonlinear activation function to get the output value of the neuron which will be sent out to neurons in the following layers. The typical architecture of a neural network is shown in Figure 2-4. If the depth of a neural network (number of hidden layers) becomes large (typically more than three hidden layers), the network is usually referred to as a *deep neural network*. T sub-field of machine learning that focuses on deep neural network studies is known as *deep learning*. The advantages of deep neural networks has been shown in a variety of fields, including image recognition, natural language processing and time series analysis.

Before starting the network training process, all weights and biases are initialized to random numbers, typically evenly distributed between -1 and 1. During training, training samples are fed into the neural network and transformed to fitted outputs via complex calculation of the neural network. Using the fitted outputs and ground truth targets, the fitting error for the training data set is evaluated via a chosen *loss function*. A chosen network optimizer like stochastic gradient descent with momentum [45] and Adam [46] then continually adjusts the learnable weights and biases based on fitting error to improve goodness of fit until the fitting errors become small enough. Since the network parameters are optimized on the training set, the trained parameters can easily overfit the training data if the training set is small and or noisy, and the capacity of network is large enough to "memorize" the training samples. To reduce overfitting, several techniques, including weight decay [47], early stopping [48]

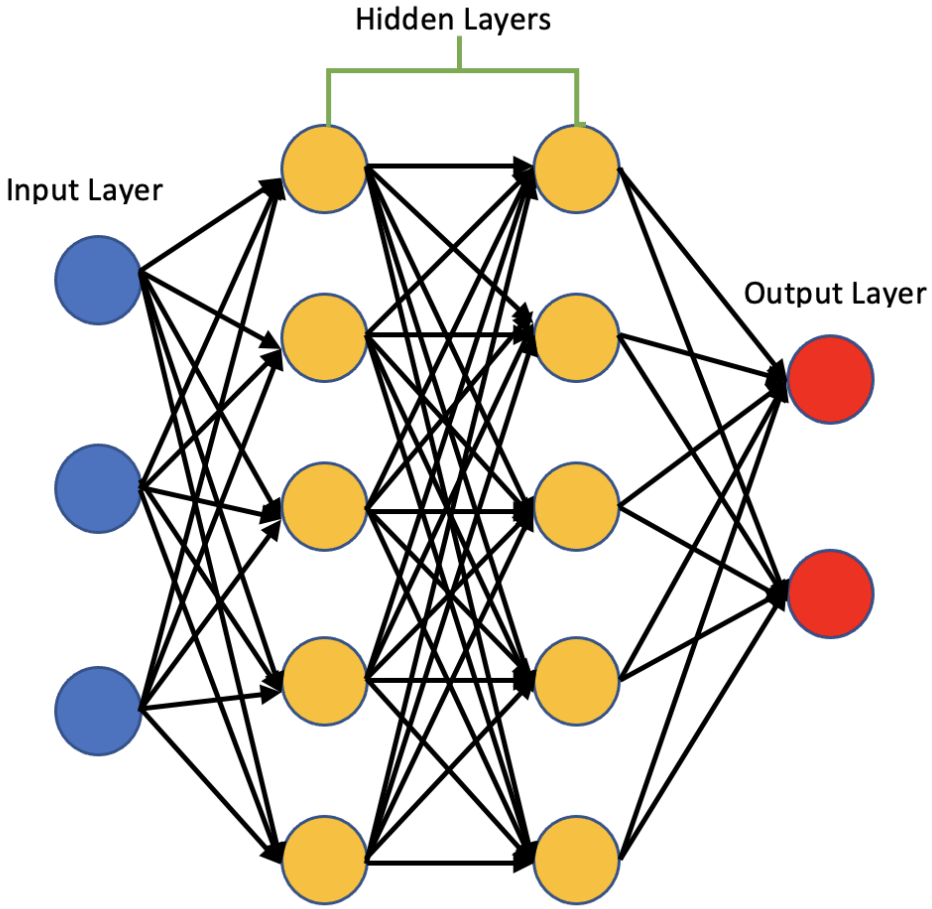

Figure 2-4: A diagram shows the architecture of a simple neural network

and dropout [49] have been developed over the past 20 years, and the availability of large labelled datasets,such as ImageNet [50], can be used to help avoid overfitting with large neural networks.

2.3 The cross-machine adaptation challenge of disruption prediction/avoidance algorithms

As mentioned in Section 1.7, near-future burning plasma tokamaks like ITER may be very vulnerable to high current and high stored energy disruptions, and the success of ITER's missions demands that a highly reliable disruption predictor be functional before the beginning of full performance operation. This requirement imposes significant challenges to the development of both physics-driven and data-driven disruption prediction algorithms. For physics-driven models, emergence of new physics on these burning-plasma machine can make the extrapolation of current models to a next-step machine very uncertain. For data-driven predictors, a disruption prediction algorithm with strong cross machine generalization ability needs to be developed and an efficient cross machine/regime adaptive strategy for training the predictor is also needed to better extract relevant knowledge from the discharges on existing machines and the early, low performance discharges on the target near-future device.

Different tokamak devices have different operational spaces, spatio-temporal scales for physics events, and plasma diagnostic sets, as indicated in Section 1.8. Therefore, most of these data-driven approaches so far were developed and optimized specifically for one device and did not show promising cross-device predictive abilities [27–29, 33, 35]. Specifically, cross-machine studies, such as [35], focused on predictors that were trained on datasets purely or mostly from one device: these predictors achieved excellent performance on the training device, but lacked the generalization capabilities derived from an understanding of the underlying physics, and therefore tended to fail on new, previously unseen device data. To overcome this difficulty, several previous studies [51, 52], explored the strategy of building a predictor from scratch. In these studies, researchers gradually add data in chronological order to retrain the predictors, and then test on future unseen discharges. However, these studies are conducted using discharges from similar operational regimes, which implicitly assumes that we can explore and learn on data that have similar parameters to future 'test' discharges. Although this assumption is generally valid for existing machines, the ITER research plan [53] suggests this will probably not be sufficient for future devices like ITER, because unmitigated disruptions in high performance (HP) regimes threaten the integrity of the facility, and we have to predict these disruptions using low performance (LP) data from these devices. A recently published work by

Kates-Harbeck et al. [37], demonstrated for the first time the potential of predictors based on deep learning (DL) for acquiring a general representation of experimental data that can be used in cross-machine applications. Therefore, given all these previous studies, developing a cross-machine DL based disruption prediction algorithm and exploring an efficient adaptive training strategy for the cross-machine prediction algorithm can be a strong candidate for realizing a sufficiently accurate disruption prediction model.

This thesis demonstrates four data-driven studies to tackle the cross-machine adaptation challenge of disruption prediction. First, a hybrid deep-learning model (discussed in chapter 3) for cross-machine disruption prediction is developed using large disruption databases from Alcator C-Mod, DIII-D and EAST. The HDL model has shown state-of-the-art performance on multiple tokamaks, with very limited machine-specific hyperparameter tuning. Second, in chapter 4, a scenario adaptive study that aims to explore an efficient training strategy for a data-driven disruption predictor is presented. The main conclusions from this study provide a possible strategy for the development of data-driven disruption predictors on next-step tokamaks. Third, the upgraded HDL model, that incorporates the predictive capability using various plasma unstable events, discussed in Section 1.5, is elaborated in chapter 5. By predicting the event chain towards final disruption, the upgraded HDL model can facilitate the investigation of disruption causes, and enable the avoidance of impending disruptions. Finally, a data-driven symbolic boundary, for predicting n=1 tearing mode (TM) onset across tokamaks, is discussed in in chapter 6, which provides an example of finding an explicit plasma instability boundary via data-driven methods.

References - Chapter 2

[1] P. C. de Vries, G. Pautasso, D. Humphreys, M. Lehnen, S. Maruyama, J. A. Snipes, A. Vergara, and L. Zabeo. Requirements for triggering the ITER disruption mitigation system. *Fusion Science and Technology*, 69(2):471–484, 2016.

[2] T.C. Hender, J.C. Wesley, J. Bialek, A. Bondeson, A.H. Boozer, R.J. Buttery, A. Garofalo, T.P. Goodman, R.S. Granetz, Y. Gribov, O. Gruber, M. Gryaznevich, G. Giruzzi, S. Günter, N. Hayashi, P. Helander, C.C. Hegna, D.F. Howell, D.A. Humphreys, G.T.A. Huysmans, A.W. Hyatt, A. Isayama, S.C. Jardin, Y. Kawano, A. Kellman, C. Kessel, H.R. Koslowski, R.J. la Haye, E. Lazzaro, Y.Q. Liu, V. Lukash, J. Manickam, S. Medvedev, V. Mertens, S.V. Mirnov, Y. Nakamura, G. Navratil, M. Okabayashi, T. Ozeki, R. Paccagnella, G. Pautasso, F. Porcelli, V.D. Pustovitov, V. Riccardo, M. Sato, O. Sauter, M.J. Schaffer, M. Shimada, P. Sonato, E. J. Strait, M. Sugihara, M. Takechi, A.D. Turnbull, E. Westerhof, D.G. Whyte, R. Yoshino, and H. Zohm. Chapter 3: Mhd stability, operational limits and disruptions. *Nuclear Fusion*, 47(6):S128–S202, June 2007.

[3] Allen H. Boozer. Theory of tokamak disruptions. *Physics of Plasmas*, 19(5), 2012.

[4] E. J. Strait, J. L. Barr, M. Baruzzo, J. W. Berkery, R. J. Buttery, P. C. De Vries, N. W. Eidietis, R. S. Granetz, J. M. Hanson, C. T. Holcomb, D. A. Humphreys, J. H. Kim, E. Kolemen, M. Kong, M. J. Lanctot, M. Lehnen, E. Lerche, N. C. Logan, M. Maraschek, M. Okabayashi, J. K. Park, A. Pau, G. Pautasso, F. M. Poli, C. Rea, S. A. Sabbagh, O. Sauter, E. Schuster, U. A. Sheikh, C. Sozzi, F. Turco, A. D. Turnbull, Z. R. Wang, W. P. Wehner, and L. Zeng. Progress in disruption prevention for ITER. *Nuclear Fusion*, 59(11):112012, 2019.

[5] Fernanda G. Rimini, Diogo Alves, Gilles Arnoux, Matteo Baruzzo, Eva Belonohy, Ivo Carvalho, Robert Felton, Emmanuel Joffrin, Peter Lomas, Paul McCullen, Andre Neto, Isabel Nunes, Cedric Reux, Adam Stephen, Daniel Valcarcel, and Sven Wiesen. The development of safe high current operation in JET-ILW. *Fusion Engineering and Design*, 96-97:165–170, 2015. Proceedings of the 28th Symposium On Fusion Technology (SOFT-28).

[6] G Pautasso, C.J Fuchs, O Gruber, C.F Maggi, M Maraschek, T Pütterich, V Rohde, C Wittmann, E Wolfrum, P Cierpka, M Beck, and the ASDEX Upgrade Team. Plasma shut-down with fast impurity puff on ASDEX Upgrade. *Nuclear Fusion*, 47(8):900–913, jul 2007.

[7] Cédric Reux, Michael Lehnen, Uron Kruezi, Stefan Jachmich, Peter Card, Kalle Heinola, Emmanuel Joffrin, Peter J. Lomas, Stefan Marsen, Guy Matthews, Valeria Riccardo, Fernanda Rimini, and Peter de Vries. Use of the disruption mitigation valve in closed loop for routine protection at JET. *Fusion Engineering and Design*, 88(6):1101–1104, 2013. Proceedings of the 27th Symposium On Fusion Technology (SOFT-27); Liège, Belgium, September 24-28, 2012.

[8] R. Sweeney, W. Choi, R.J. La Haye, S. Mao, K.E.J. Olofsson, and F.A. Volpe. Statistical analysis of m / n = 2/1 locked and quasi-stationary modes with rotating precursors at DIII-D. *Nuclear Fusion*, 57(1):016019, Jan 2017.

[9] Vitus Mertens, Gerhard Raupp, and Wolfgang Treutterer. Chapter 3: Plasma Control in ASDEX Upgrade. *Fusion Science and Technology*, 44(3):593–604, 2003.

[10] N.W. Eidietis, W. Choi, S.H. Hahn, D.A. Humphreys, B.S. Sammuli, and M.L. Walker. Implementing a finite-state off-normal and fault response system for disruption avoidance in tokamaks. *Nuclear Fusion*, 58(5):056023, mar 2018.

[11] N.M. Trang Vu, T.C. Blanken, F. Felici, C. Galperti, M. Kong, E. Maljaars, and O. Sauter. Tokamak-agnostic actuator management for multi-task integrated control with application to TCV and ITER. *Fusion Engineering and Design*, 147:111260, 2019.

[12] W. Treutterer, R. Cole, K. Lüddecke, G. Neu, C. Rapson, G. Raupp, D. Zasche, and T. Zehetbauer. ASDEX Upgrade discharge control system—a real-time plasma control framework. *Fusion Engineering and Design*, 89(3):146–154, 2014. Design and implementation of real-time systems for magnetic confined fusion devices.

[13] S. A. Sabbagh, J. W. Berkery, Y. S. Park, J. H. Ahn, J. Bialek, Y. Jiang, J. D. Riquezes, J. G. Bak, S. H. Hahn, J. Kim, J. Ko, J. Lee, S. W. Yoon, C. Ham, A. Kirk, L. Kogan, D. Ryan, A. Thornton, M. Boyer, K. Erickson, Z. Wang, V. Klevarova, and G. Pautasso. Disruption Event Characterization and Forecasting in Tokamaks and Expansion to Real-Time Application*. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP17.004, Jan 2020.

[14] J. W. Berkery, S. A. Sabbagh, R. E. Bell, S. P. Gerhardt, and B. P. LeBlanc. A reduced resistive wall mode kinetic stability model for disruption forecasting. *Physics of Plasmas*, 24(5):056103, 2017.

[15] J. D. Riquezes, S. A. Sabbagh, J. W. Berkery, Y. S. Park, J. H. Ahn, Y. Jiang, J. Butt, E. Fredrickson, and J. G. Bak. Rotating MHD Mode Analysis Including Real-time data on KSTAR Supporting Disruption Event Characterization and Forecasting. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP17.007, Jan 2020.

[16] J. Butt, S. A. Sabbagh, Y. S. Park, J. H. Ahn, J. W. Berkery, Y. Jiang, and J. D. Riquezes. ELM Detection Capability for Disruption Event Characterization and Forecasting. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP17.006, January 2020.

[17] Won Ha Ko, SW Yoon, WC Kim, JG Kwak, KR Park, YU Nam, SJ Wang, J Chung, BH Park, GY Park, et al. Kstar overview. *Bulletin of the American Physical Society*, 2022.

[18] M. Maraschek, A. Gude, V. Igochine, H. Zohm, E. Alessi, M. Bernert, C. Cianfarani, S. Coda, B. Duval, B. Esposito, S. Fietz, M. Fontana, C. Galperti, L. Giannone, T. Goodman, G. Granucci, L. Marelli, S. Novak, R. Paccagnella, G. Pautasso, P. Piovesan, L. Porte, S. Potzel, C. Rapson, M. Reich, O. Sauter, U. Sheikh, C. Sozzi, G. Spizzo, J. Stober, and W. Treutterer. Path-oriented early reaction to approaching disruptions in ASDEX Upgrade and TCV in view of the future needs for ITER and DEMO. *Plasma Physics and Controlled Fusion*, 60(1):14047, 2018.

[19] U.A. Sheikh, B.P. Duval, C. Galperti, M. Maraschek, O. Sauter, C. Sozzi, G. Granucci, M. Kong, B. Labit, A. Merle, N. Rispoli, and and. Disruption avoidance through the prevention of NTM destabilization in TCV. *Nuclear Fusion*, 58(10):106026, Aug 2018.

[20] D. Humphreys, A. Kupresanin, M. D. Boyer, J. Canik, C. S. Chang, E. C. Cyr, R. Granetz, J. Hittinger, E. Kolemen, E. Lawrence, V. Pascucci, A. Patra, and D. Schissel. Advancing fusion with machine learning research needs workshop report. *Journal of Fusion Energy*, 39:123–155, 2020.

[21] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 3 edition, 2006.

[22] Max Kuhn and Kjell Johnson. *Feature engineering and selection: A practical approach for predictive models*. CRC Press, 2019.

[23] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[24] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, second edition, 2009.

[25] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

[26] Akanksh Basavaraju, Jing Du, Fujie Zhou, and Jim Ji. A machine learning approach to road surface anomaly assessment using smartphone sensors. *IEEE Sensors Journal*, 20(5):2635–2647, 2020.

[27] K. J. Montes, C. Rea, R. S. Granetz, R. A. Tinguely, N. Eidietis, O. M. Meneghini, D. L. Chen, B. Shen, B. J. Xiao, K. Erickson, and M. D. Boyer. Machine learning for disruption warnings on Alcator C-Mod, DIII-D, and EAST. *Nuclear Fusion*, 59(9):096015, 2019.

[28] C. Rea, K. J. Montes, K. G. Erickson, R. S. Granetz, and R. A. Tinguely. A real-time machine learning-based disruption predictor in DIII-D. *Nuclear Fusion*, 59(9):096016, 2019.

[29] Yichen Fu, David Eldon, Keith Erickson, Kornee Kleijwegt, Leonard Lupin-Jimenez, Mark D. Boyer, Nick Eidietis, Nathaniel Barbour, Olivier Izacard, and Egemen Kolemen. Machine learning control for disruption and tearing mode avoidance. *Physics of Plasmas*, 27(2):022501, 2020.

[30] Jesús Vega, Sebastián Dormido-Canto, Juan M. López, Andrea Murari, Jesús M. Ramírez, Raúl Moreno, Mariano Ruiz, Diogo Alves, and Robert Felton. Results of the JET real-time disruption predictor in the ITER-like wall campaigns. *Fusion Engineering and Design*, 88(6-8):1228–1231, Oct 2013.

[31] G. A. Rattá, J. Vega, and A. Murari. Viability Assessment of a Cross-Tokamak AUG-JET Disruption Predictor. *Fusion Science and Technology*, 74(1-2):13–22, 8 2018.

[32] A. Murari, M. Lungaroni, E. Peluso, P. Gaudio, J. Vega, S. Dormido-Canto, M. Baruzzo, and M. Gelfusa. Adaptive predictors based on probabilistic SVM for real time disruption mitigation on JET. *Nuclear Fusion*, 58(5):056002, 5 2018.

[33] R. Aledda, B. Cannas, A. Fanni, A. Pau, and G. Sias. Improvements in disruption

prediction at ASDEX Upgrade. *Fusion Engineering and Design*, 96-97:698–702, 10 2015.

[34] W. Zheng, F.R. Hu, M. Zhang, Z.Y. Chen, X.Q. Zhao, X.L. Wang, P. Shi, X.L. Zhang, X.Q. Zhang, Y.N. Zhou, Y.N. Wei, and Y. Pan. Hybrid neural network for density limit disruption prediction and avoidance on J-TEXT tokamak. *Nuclear Fusion*, 58(5):056016, May 2018.

[35] C.G Windsor, G. Pautasso, C. Tichmann, R.J Buttery, T.C Hender, and the ASDEX Upgrade Contributors, JET EFDA and Team. A cross-tokamak neural network disruption predictor for the JET and ASDEX Upgrade tokamaks. *Nuclear Fusion*, 45(5):337, May 2005.

[36] A. Piccione, J. W. Berkery, S. A. Sabbagh, and Y. Andreopoulos. Physics-guided machine learning approaches to predict the ideal stability properties of fusion plasmas. *Nuclear Fusion*, 60(4):046033, 2020.

[37] Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568:526–531, 2019.

[38] J. Zhu, C. Rea, K. J. Montes, R. Granetz, R. Sweeney, and R. A. Tinguely. Hybrid deep learning architecture for general disruption prediction across tokamaks. *Nuclear Fusion*, 61(2):026007, 2020.

[39] Diogo R. Ferreira, Pedro J. Carvalho, Carlo Sozzi, Peter J. Lomas, and JET Contributors. Deep learning for the analysis of disruption precursors based on plasma tomography. *Fusion Science and Technology*, 76(8):901–911, 2020.

[40] R. M. Churchill, B. Tobias, and Y. Zhu. 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*, 27(6):062510, 2020.

[41] A. Pau, A. Fanni, S. Carcangiu, B. Cannas, G. Sias, A. Murari, and F. Rimini. A machine learning approach based on generative topographic mapping for disruption prevention and avoidance at JET. *Nuclear Fusion*, 59(10):106017, 2019.

[42] A. Murari, J. Vega, G.A. Rattá, G. Vagliasindi, M.F. Johnson, and S.H. Hong. Unbiased and non-supervised learning methods for disruption prediction at JET. *Nuclear Fusion*, 49(5):055028, 5 2009.

[43] Raffaele Aledda, Barbara Cannas, Alessandra Fanni, Giuliana Sias, and Gabriella Pautasso. Adaptive mapping of the plasma operational space of ASDEX Upgrade for disruption prediction. *International Journal of Applied Electromagnetics and Mechanics*, 39(1-4):43–49, 2012.

[44] Y. Wei, J. W. Brooks, R. Chandra, J. P. Levesque, Boting Li, A. Saperstein, I. G. Stewart, M. E. Mauel, G. A. Navratil, and C. Hansen. A dimensionality reduction algorithm for mapping tokamak operation regimes using variational autoencoder (VAE) neural network. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP16.007, Jan 2020.

[45] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[46] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization.

arXiv preprint arXiv:1412.6980, 2014.

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[48] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[49] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[51] S. Dormido-Canto, J. Vega, J.M. Ramírez, A. Murari, R. Moreno, J.M. López, and A. Pereira. Development of an efficient real-time disruption predictor from scratch on JET and implications for ITER. *Nuclear Fusion*, 53(11):113001, 11 2013.

[52] J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, A. Pereira, and A. Acero. Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks. *Nuclear Fusion*, 54(12):123001, Dec 2014.

[53] DJ Campbell et al. The iter research plan. In *Proceedings of the 24th International Conference on Fusion Energy, San Diego, CA, USA*, pages 8–13, 2012.

Chapter 3

Hybrid Deep-Learning Disruption Prediction Model

The cross-machine disruption prediction studies in this thesis are conducted using the hybrid deep-learning predictor, a machine learning algorithm based on a deep neural network (an introduction of neural network model can be found in Section 2.2.3). The abbreviation HDL, for hybrid deep-learning, is used throughout this thesis. This chapter describes the design, architecture and test processes of the HDL model in detail and summarizes the current multi-machine performance of HDL model and the basic cross-machine numerical experiments using the HDL model.

Section 3.1 gives an introduction to deep neural networks (DNN) and several commonly used DNN layers. After this, the disruption warning databases [1–4] that are used to implement the data-driven studies throughout this thesis are described in Section 3.2, and the datasets extracted from disruption warning databases for studies in this chapter are discussed in detail. Following the description of the datasets, application of an unsupervised dimensionality-reduction algorithm to the high-dimensional plasma data is presented in Section 3.3. Based on several important findings from the explorative data analysis in Section 3.3, a new HDL disruption prediction model is developed to yield improved learning from the temporal plasma signals. The design of the HDL model is presented in Section 3.4, and the performance comparison between the HDL model and several other data-driven disruption prediction models is also discussed in Section 3.4. An extensive cross-machine numerical disruption prediction study, based on the HDL model is elaborated in Section 3.5. This includes the motivation, design and results of the cross-machine numerical experiments, as well as several important qualitative conclusions drawn from the results that can inform the development of data-driven disruption prediction for future devices. Finally, a summary of the HDL model development and the cross-machine experimental results are given in Section 3.6.

3.1 Deep Neural Network Model

As mentioned in Section 2.2.3, a neural network is a computational model that consists of several interconnected layers of artificial neurons. Although the universal approximation theorems [5, 6] imply that a feed-forward neural network, with a single arbitrarily wide hidden layer, and non-polynomial activation, can sufficiently fit any Borel measurable function from one finite-dimensional space to another. However, such a single layer might be infeasibly large, and thus the training algorithm could be unable to find the values of the parameters that correspond to the function we need. In many cases, choosing a deeper neural network model can greatly reduce the number of neurons required to fit the desired function, and hence decrease the generalization error [7]. This empirical observation provides the motivation for deep neural network studies. The deep neural network model is a subset of the neural network model family, usually having more than three hidden layers. Given the availability of large labelled databases and powerful computing systems, the deep neural network model has been shown to significantly outperform other machine learning models in a variety of tasks, including computer vision [8], natural language processing [9] and recommendation systems [10]. A basic type of deep neural network, the fully connected (FC) neural network, has already been introduced in Section 2.2.3. Two other commonly used types of deep neural network, the convolutional neural network (CNN) and recurrent neural network (RNN), are discussed in this section.

3.1.1 Convolutional Neural Network

The CNN is a class of neural network model that is typically used in computer vision tasks. The key structure of a CNN, the convolutional layer, is a regularized version of an FC layer with shared weight across many neurons; it is based on several convolution kernels or filters that move along certain axes of the input data, giving translation-equivariant outputs (i.e. a translation of input features results in an equivalent translation of outputs). Given an input data matrix, the convolutional layer performs a Frobenius inner product of the convolution kernel with the input matrix and then transforms the resultant product using a nonlinear activation function. One commonly used activation is the rectified linear unit, ReLU [11]. As the kernel with activation function slides along the pre-set directions of the input matrix, it generates one channel of the output feature map. The outputs from different kernels of a convolutional layer are stacked to form the final output feature map. A simple diagram that explains the basic logic of the convolutional layer is shown in Figure 3-1.

A typical convolutional neural network has three main types of layers: a con-

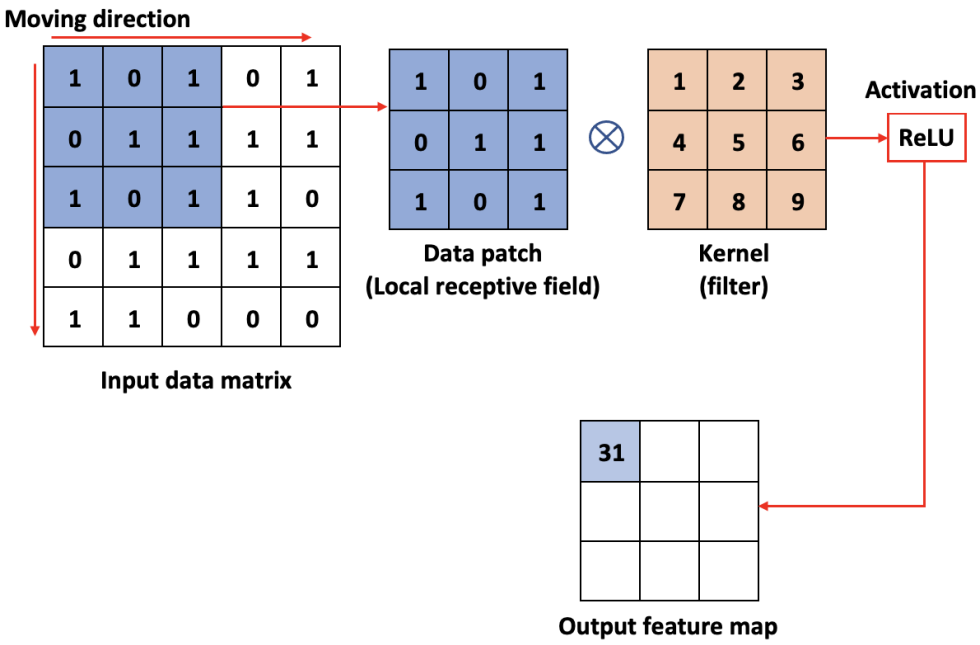

Figure 3-1: Simple diagram explains the basic operation of the convolutional layer

volutional layer, a pooling layer and a fully connected layer.The convolutional layer, shown in Figure 3-1, is the key component of the CNN, and is usually the first layer of the whole network. The pooling layer, also called the down-sampled layer,can reduce the number of input parameters by sweeping a kernel/filter across the entire input matrix. Unlike the convolutional layer, the pooling kernel does not have any free parameters; it applies an aggregation function to the data in the local receptive field, and generates a "summary" of the data patch. There are two main types of pooling kernels: 1) a max pooling kernel that outputs the maximum within the receptive field; and 2) an average pooling kernel that outputs the average value over the receptive field. Although some information might be lost through the pooling operation, it helps reduce the complexity and hence reduces the risk of overfitting. The fully connected layers that are usually the final layers of the CNN can do the classification or regression tasks to generate the final output of the CNN. A deep CNN can consist of more than a hundred layers. With each layer, the CNN increases in complexity, allowing it to detect more abstract and more complex features, and identify greater portions of the input data. A typical deep CNN architecture is shown in Figure 3-2 [8].

3.1.2 Recurrent Neural Network

A recurrent neural network is a type of feed-forward neural network that is specifically designed for sequence data processing. The connections between nodes of an

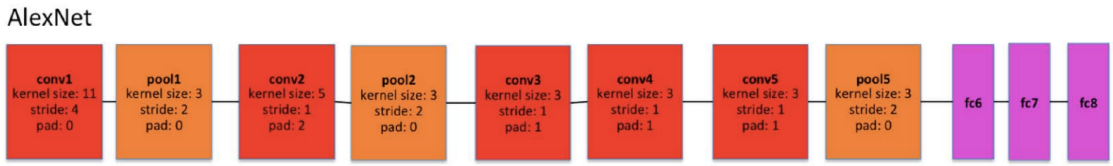

Figure 3-2: Simple illustration of the architecture of AlexNet [8] from [12]

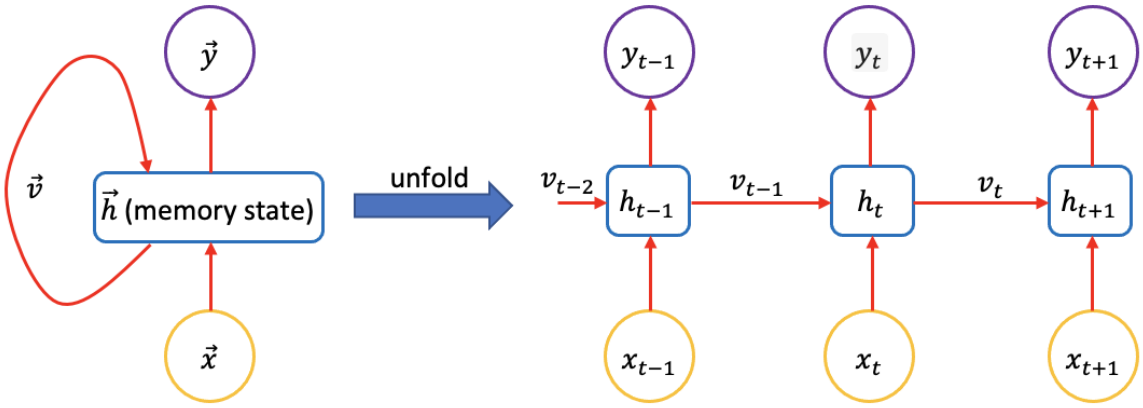

Figure 3-3: Simple illustration of the architecture of basic RNNs. In the left part, $\vec{x}$, $\vec{y}$, $\vec{h}$, $\vec{v}$ represent the input sequence, output sequence, hidden state sequence and information flowing between consecutive time steps respectively. In the right part, a individual hidden layer of RNNs is unfolded to explain the mechanism of the neural network

RNN form a graph along input sequences. The key difference between RNNs and CNNs/dense NNs is that RNNs maintain a hidden state, called memory, which is calculated from previous inputs. The current outputs of RNNs are derived from both the hidden state and the current inputs. A simple diagram explaining the mechanism of an RNN is given in Figure 3-3. The left part shows the compressed/folded visual of the RNN that represents the whole neural network, and the right part shows the unfolded visual of the RNN that represents the mechanism of individual RNN layers.

Using the vanilla RNN shown in Figure 3-3, it may be difficult or even impossible to process the information from the early time steps of the input sequences. When we backpropagate gradients through layers and also through time, we need to sum up all the previous contributions until the current one which will introduce the product of the partial derivative of the hidden state to the gradients and this product can easily goes to 0 or infinity. From the computational perspective, this means that when training a basic RNN, using stochastic gradient descent, the long term gradients usually "vanish" (go to zero) or "explode" (go to infinity) as a result of the recurrent process shown in Figure 3-3 [13, 14]. To tackle this problem, two commonly used

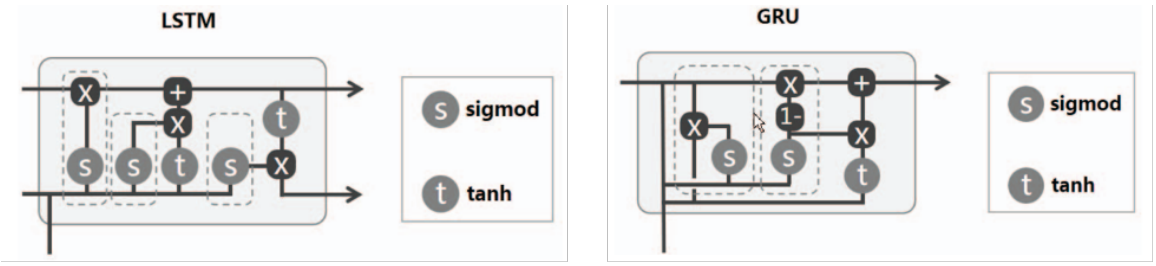

Figure 3-4: Simple illustration of the architecture of LSTM cell (left) and GRU cell (right) from [17]

variants of the basic RNN, long short-term memory (LSTM) and gated recurrent unit (GRU) were introduced in 1997 [15] and 2014 [16] respectively. LSTMs modify the simple hidden states to "cells" with three gates—an input gate, an output gate, and a forget gate, each to control the information flow. These gates allow gradients to flow unchanged. GRUs are very similar to LSTMs, also using gates to control the information flow. The difference between LSTMs and GRUs is that GRUs do not have additional "cell" states, but instead have two gates—a reset gate and an update gate instead of the three gates found in LSTMs. The schematic structures of LSTM and GRU nodesare shown in Figure 3-4.

3.2 Dataset Description

The disruption prediction studies in this chapter are conducted on disruption warning datasets from three machines [4]: Alcator C-Mod (2012-2016 campaigns), DIII-D (2014-2018 campaigns) and EAST (2014-2018 campaigns). For all three datasets, we include all types of disruptions except for intentional ones (planned disruptions for disruption physics study). The choice of which parameters to include in the databases is guided by our knowledge of the plasma physics mechanisms inherent to disruption characteristics of the different devices, as well as the availibility and consistency of these parameters for all three machines. Our choices for many of the disruption-relevant parameters included in this study are also influenced by several previous papers and investigations [4, 18, 19]. The signals considered for the predictive models reported in this thesis, and their definitions, can be found in Table 3.1, while the composition of the three training datasets is shown in Table 1.1. Given these databases, we formalize the disruption prediction problem in a *sequence-to-label* supervised machine learning framework, where we assign a label to each input plasma sequence, S (a 10-step consecutive sequence in time of 12 plasma signals) and train an algorithm to learn the functional representation. The input sequences are then mapped to one of two possible labels, 'disruptive' or 'non-disruptive'. To this aim, we

Table 3.1: Descriptions and symbols of all considered signals [22]

Signal description	Symbol
$\frac{\textbf{Plasma current}-\textbf{programmed plasma current}}{\textbf{Programmed plasma current}}$	`ip-error-fraction`
Perturbed field of nonrotating mode[a], $\frac{\mathbf{B^{n=1}}}{\mathbf{B_{tor}}}$	`locked-mode-proxy`
$\frac{\textbf{Electron density}}{\textbf{Greenwald density}}$	`Greenwald-fraction`
Distance between the plasma and the lower divertor	`lower-gap`
Current centroid vertical position error[b]	`z-error-proxy`
Plasma elongation	`kappa`
Poloidal beta	`betap`
$\frac{\textbf{Radiated power}}{\textbf{Input power}}$	`radiated-fraction`
Standard deviation of the magnetic field[c] **measured from an array of Mirnov coils, normalized by $\mathbf{B_{tor}}$**	`rotating-mode-proxy`
Loop Voltage $\mathbf{V_{loop}}$	`v-loop`
Normalized internal inductance	`li`
Safety factor at 95% flux surface	`q95`

[a]For the C-Mod database, the `locked-mode-proxy` signal is obtained from a Mirnov coil array instead of the saddle coil.
[b]For the DIII-D database, we use current centroid vertical position instead of position error for the `z-error-proxy signal`.
[c]For the DIII-D database, we use n=1 component of magnetic field measured from a Mirnov coil array normalized by B_{tor} for the `rotating-mode-proxy` signal.

explicitly defined different time thresholds for each machine to identify the unstable phase of the disruptive training discharges and assigned the disruptive label to plasma sequences that intersect the unstable phase of disruptive experimental runs, while the non-disruptive label is assigned to sequences extracted from the non-disruptive discharges. This classification scheme implicitly assumes that it is possible to detect a transition in time from a safe operational regime to a disruptive one, and is another instance of incorporating physics knowledge into the AI workflow [20, 21]. The chosen time thresholds vary for the three devices, depending on the transition points where some of the plasma parameters exhibit identifiable changes in behavior for a notable fraction of disruptions before disruptions occurring [4] and the suggestions from tokamak operators.

The training samples are ordered into sequences of ten time slices extracted from each shot of the training dataset. For each shot, we randomly select a subset of

Table 3.2: The dataset composition of the three disruption warning databases [22]

Device	No. training shots	No. test shots	No. validation shots	Sampling rate (ms)	Time Threshold (ms)	No. samples per training shot
C-Mod	3343 (692)	651 (142)	463 (98)	5	75	16
DIII-D	5286 (732)	1085 (157)	734 (107)	10	400	25
EAST	8296 (2301)	1674 (475)	1137 (322)	25	500	20

Values in parentheses give the number of disruptive shots within each dataset.

examples: this is one of the model's hyperparameters, tuned for each machine. The disruptive training sequences are randomly extracted from all sequences that intersect the unstable phase of each disruptive shot, while those sequences outside of the unstable region are not included in the training set. If disruptive patterns are learned properly, the algorithm will also be able to identify similar trends at times prior to the formally set time threshold, enabling the detection of early disruptive precursors. The non-disruptive sequences are randomly extracted from the flattop of non-disruptive training discharges. It is interesting to note that the database population consists of mostly non-disruptive data, thus resulting in a dataset imbalanced with respect to disruptive data which can hamper the training of the disruption predictor.

3.3 Explorative data analysis through an unsupervised learning algorithm

Disruptions are typically well characterized by high-dimensional data from multiple plasma signals, which complicates both analysis and physics interpretation when studying disruptive events. In this section, we discuss the application of a nonlinear dimensionality reduction technique called t-Distributed Stochastic Neighbor Embedding (t-SNE) [23] (see appendix A.2 of [22] for details of the t-SNE algorithm) to visualize high-dimensional plasma data in a 2-D plane to study the inherent data structure of the considered plasma signals. In principle, the t-SNE algorithm can be applied to any high-dimensional database. However, in this section, we only show the application to the C-Mod database, as it is considered more difficult to predict through a data-driven approach than is the case for EAST and DIII-D [1]. The analysis of the DIII-D and EAST disruption databases can be found in [22].

Figure 3-5 shows the t-SNE algorithm applied to time slice data (left) and aggregated sequence data (right) for the C-Mod disruption warning database. In the left subplot, each blue point represents a randomly sampled time slice (a 12-element array composed of the 12 plasma signals from Table 3.1) taken from the flattop of a

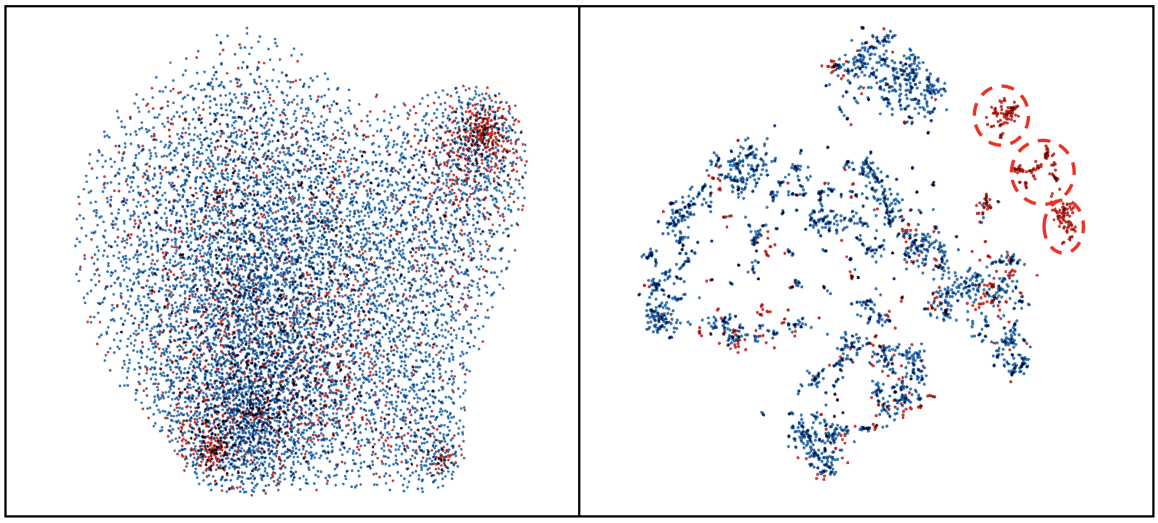

Figure 3-5: t-SNE clustering for visualizing C-Mod data. On the left, t-SNE is performed on individual disruptive (red) and non-disruptive (blue) time slices. On the right, t-SNE is performed on 10-step disruptive (red) and non-disruptive (blue) sequences. Three major clusters of disruptive data can be isolated (as shown by the dashed circles). The colouring is done a-posteriori.

non-disruptive shot, while each red point represents a time slice randomly sampled from the last 75 ms of a disruptive shot. On the right, each red point represents a 10-step (a 10×12 element matrix) sequence randomly sampled from the last 75 ms of a disruptive sho, while each blue point represents a 10-step sequence randomly sampled from the flattop of a non-disruptive shot. We include all disruptions, without discriminating by the cause. The coloring of each data point in the plots is done *a-posteriori*, i.e. it was not provided during the training process, thereby characterizing the t-SNE as an unsupervised clustering technique.

Several important conclusions can be drawn. First, the clustering of individual time slices does not isolate clear data clusters in the low-dimensional map. However, by performing t-SNE on 10-step plasma sequences, it is possible to isolate three major clusters - identified by dashed circles in Figure 3-5 - which account for approximately 60% of the disruptive data. The improved separation of disruptive and non-disruptive data, obtained when clustering sequences, highlights the importance of temporal correlation and mutual information among consecutive time slices. This further suggests that sequence-based classifiers could have a clear advantage over the single time slice based ones. Secondly, the t-SNE application to C-Mod sequences reveals that a substantial fraction ($\approx 40\%$) of disruptive sequences remains mixed with non-disruptive sequences. Further analysis of such data finds that these disruptive sequences are primarily linked to fast radiative collapses caused by molybdenum impurities. These disruptions have very short warning time between first identifiable disruption pre-

cursor and disruption onset, up to a few tens of milliseconds, and we argue that any data-driven disruption prediction algorithm for C-Mod would struggle to predict such cases (at least with the current set of input features) and thus be affected by a high degree of false negatives (missed predictions). The three isolated clusters, identified by red dashed circles in Figure 3-5, are representative of specific disruption dynamics, including VDEs, impurity accumulations and MHD-driven disruptions. These precursors can be identified through inspection of the specific time series. In addition, the physics insights obtained from t-SNE clustering results suggest that t-SNE can be used as interpretable "by design" method for analysis of plasma signals, allowing inspection of complex data structure and patterns and thus obtaining more thorough physics understanding

3.4 The hybrid deep-learning (HDL) disruption-prediction model

Based on our findings about the importance of temporal information, we introduced a Hybrid Deep Learning (HDL) network for time series processing. Figure 3-6(a) shows the architecture of the network used for the cross-machine disruption prediction application reported in this thesis. The HDL network consists of two GRU layers [24], one fully connected layer and three novel Multi-Scale Temporal Convolution (MSTConv) layers, plus the input and classification layers. The MSTConv layer is inspired from work in machine translation [25], and the detailed structure of one MSTConv layer is shown in Figure 3-6(b). It consists of six 1-D causal convolution layers [26] with different window lengths, L, from one to six. The first 1-D convolution layer can only access the current time step, t_0. The L^{th} 1-D convolution layer can look at L time steps from t_{0-L+1} to t_0. This structure enables different 1-D convolution layers to capture local temporal information at different levels (e.g., 1st order time derivative, 2nd order time derivative, ...). The resulting outputs from these six layers are concatenated and then processed through a batch normalization layer [27] and a ReLU (Rectified Linear Unit) activation to develop new features. It is important to highlight that different parts of the HDL architecture serve different purposes. The first two MSTConv layers are used to extract local temporal patterns from the input plasma sequences to form a richer representation of the input space. The following two GRU layers – with their long-term memory capability - can capture the long-range dependencies across different signals in the sequences. Then the subsequent MSTConv and fully connected layers can compress and summarize the output representation from the GRU layers to a 12-dimension latent encoding (dimension of the latent encoding is a tunable parameter) which can be mapped to the output by the

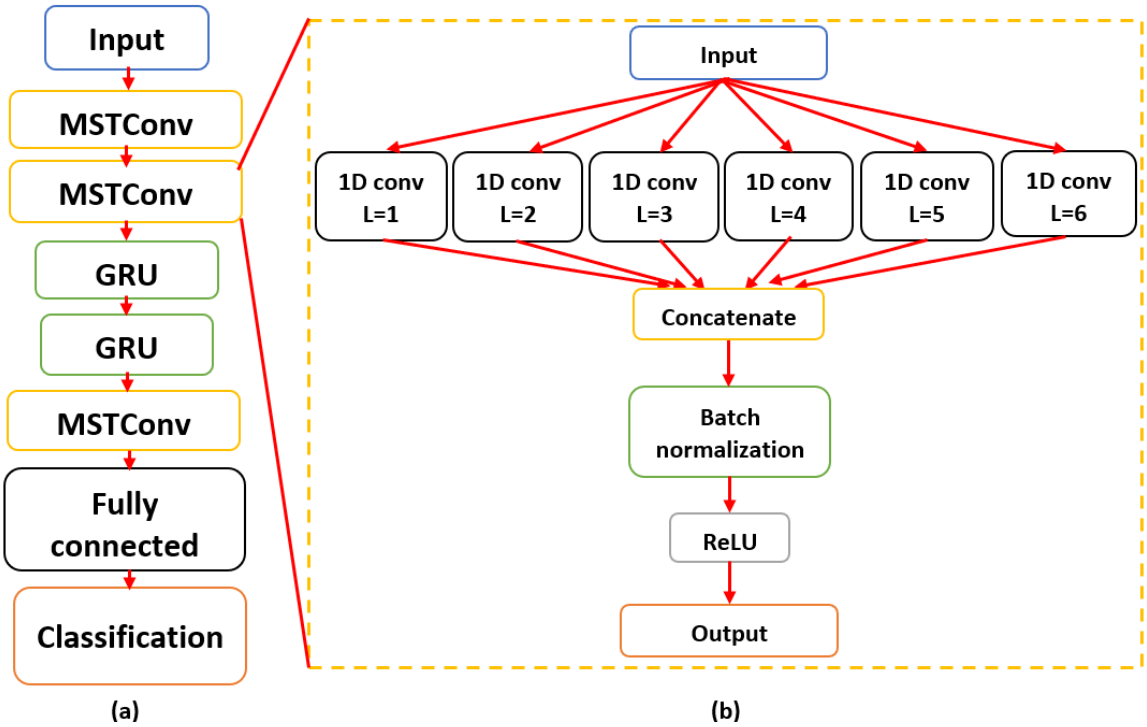

Figure 3-6: The HDL architecture (a) and the detailed structure of MSTConv layer (b). Notice that the MSTConv layer consists of 6 1-D causal convolution layers with window lengths L from 1 to 6.

classification layer.

A shot-by-shot testing scheme was designed following [4] to simulate alarms triggered in the Plasma Control System (PCS) using test shots from different devices. Given an input plasma sequence S which is a 2D matrix consisting of 10 time steps and 12 input features (i.e. a 10×12 matrix), the predictor maps S to a 'disruptivity' score between 0 and 1 at the last time step of the sequence; here, 1 is the disruptive class and 0 is the non-disruptive class. During testing, the whole flattop phase of each test shot is subdivided in batches of 10 step sequences, given the HDL architecture design. Each neighbouring testing sequence will have 9 overlapping steps, and there are N-9 sequences for a test shot with N steps. If the disruptivity exceeds a pre-set threshold – e.g., 0.7 - at any test time step, the test shot is classified as disruptive and the warning time is recorded for truly disruptive shots, defined as the difference between the alarm time and the final current quench (t_{dis}). A successfully detected disruption on C-Mod is shown in Figure 3-7: under a binary classification scheme, this is regarded as true positive (TP), while false positives (FP) correspond to a false warning, or a healthy plasma being declared to be disruptive. This latter situation can lead to premature plasma termination, but on the other hand, failure to predict a disruption early enough (false negative, FN) is even more costly, be-

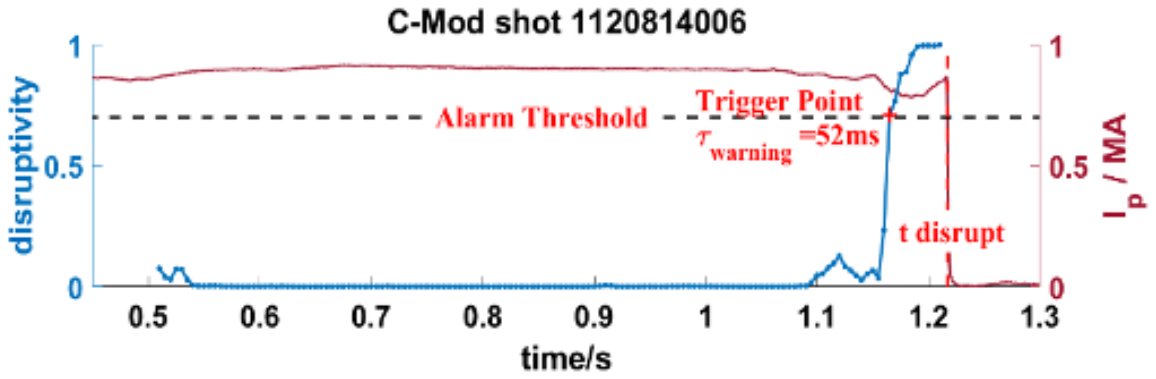

Figure 3-7: A successfully detected disruption on C-Mod.

cause it prevents any damage control of disruption consequences. A trade-off can be achieved by adjusting the alarm threshold of the disruptivity, as visually demonstrated by a receiver–operator characteristic (ROC) curve [28]. The area under the ROC curve (AUC) is used as performance metric for the HDL predictor. Throughout this thesis, we evaluate the predictive performances on all tokamaks at 50 ms before the disruption event; this is chosen as the expected minimum warning time required to successfully trigger disruption mitigation systems on future devices [29].

3.4.1 Training technicalities for the HDL model

Effective training of complex deep neural networks (NN) is a challenging task that involves several technical details, as described in [30]. Among other things, it is important to address the proper input feature normalization, and to understand which tunable parameters can increase the transferrability of the cross machine predictor while stabilizing its performance. In the following subsections, we will describe the methods implemented to tackle these challenges for optimally training our deep NN predictor.

Normalization

NNs usually need all input features to have similar numerical ranges for all training examples [30, 31]. This makes the use of raw plasma signals as inputs to any NN numerically difficult, as different signals have values that can range over many different orders of magnitude. To deal with this, all 12 signals should be normalized before being used in the network. The normalization should ideally be a common transformation such that it maps a set of signals with the same physical value from different devices to similar numerical values. Different tokamak devices have different operational spaces, spatiotemporal scales, and diagnostics. Moreover, different machines have different event chains in the lead-up to disruptions, and the most important

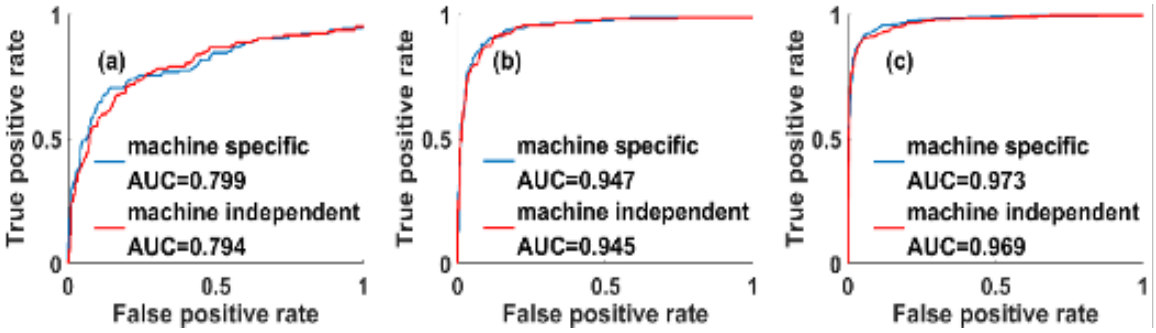

Figure 3-8: The ROC curves from test sets for machine specific normalization (blue) and machine independent normalization (red), for C-Mod (a), DIII-D (b), and EAST (c).

disruption-relevant physics parameters can be different for each machine. Therefore, such a physics-based common transformation is difficult to find, and its extrapolation to ITER is uncertain. However, we find that the best-performing method is to standardize each signal on one machine by its mean and standard deviation across the entire dataset. For each signal on one machine, its normalized form is obtained as follows: $x_{norm} = (x - \mathbf{mean}(x))/\mathbf{std}(x)$. The comprehensive list of normalization parameters can be found in appendix A.4 of [22].

The normalization process is independently applied to the data from each of the three machines, which implies it is machine-specific: this simple normalization scheme is instead chosen to solve the numerical challenge, leaving the generalized signal transformation to be done by the NN. A machine-independent normalization method has also been tested for the three datasets: this normalization standardizes all datasets with a common set of parameters. A performance comparison of HDL predictors using the two normalizations (machine-specific and machine-independent) is shown in Figure 3-8. For the machine specific cases (blue curves), the HDL predictor is trained and evaluated using the training and test sets of each machine normalized by corresponding normalization parameters. For the machine independent cases (red curves), the HDL predictor is trained and evaluated using training and test sets of each machine but normalized by 'common' normalization parameters (fixed for all 3 machines), and they give only slightly worse results, implying that the HDL performance is only weakly dependent on the normalization parameters, as long as all signals have proper numerical ranges (approximately -1 to 1).

Cross Machine Label Smoothing (CMLS)

To train a multi-machine predictor, we combine training data from different machines to form a new training set. However, direct mixing of data from various devices can result in a problem: the initial assigned target labels for other devices might not be

suitable for the new test device. For example, a certain disruptive sequence from EAST might not be that disruptive to C-Mod when compared to C-Mod's disruptive data. Also, a non-disruptive sequence from C-Mod might be slightly unstable to EAST when compared to EAST's non-disruptive data. In other words, we need to take into account the uncertainty associated with running a discharge on EAST or on DIII-D with similar operational parameters as C-Mod discharges and vice versa. To deal with this problem, we choose two smoothing parameters ϵ_1, ϵ_2 for each device (ϵ_1 for non-disruptive examples and ϵ_2 for disruptive examples) and use these two parameters to modify the target value of the training examples from multiple other machines. When we train the HDL predictor with some data from other machines, instead of using their initial (0, 1) target values for non-disruptive and disruptive examples, we modify their target values as (ϵ_1, $1 - \epsilon_2$). The new target values for non-disruptive examples are ϵ_1, and the target values for disruptive examples are $1 - \epsilon_2$. Notice that this modification is only applied to those training examples from other devices; those examples from the test device itself are not modified. We refer to this changing of the ground-truth target values as the cross machine label smoothing technique, and find that it further improves the cross-machine ability of the HDL predictor (see Table 3.3).

Hyperparameter tuning and neural networks ensemble

The HDL disruption predictor has fourteen architectural and two labeling hyperparameters for each device. Guided by our previous numerical experiments on the C-Mod dataset, we roughly scanned the hyperparameter space using a random search for all three machines' data until finding a plateau where any hyperparameter set in this region gives high performance for all three devices. Within this region, changes in hyperparameter choice will only result in minor changes to the model's performance for all three devices. Outside this region, performance on at least one device drops drastically. The hyperparameters of the HDL predictor are therefore selected from the middle of this plateau, and all following qualitative cross-machine conclusions consistently hold for all hyperparameter sets existing in this region. Additionally, our approach includes the adoption of an ensemble of twelve NNs, each one identical in their HDL architecture and tunable hyperparameters but with different initialization seeds. The final prediction comes therefore from an ensemble average. This method is well known in the ML community and has been shown to significantly improve the accuracy and stability of the predictor [32–34]. A comprehensive list of tunable hyperparameters for our HDL model can be found in appendix A.5 of [22].

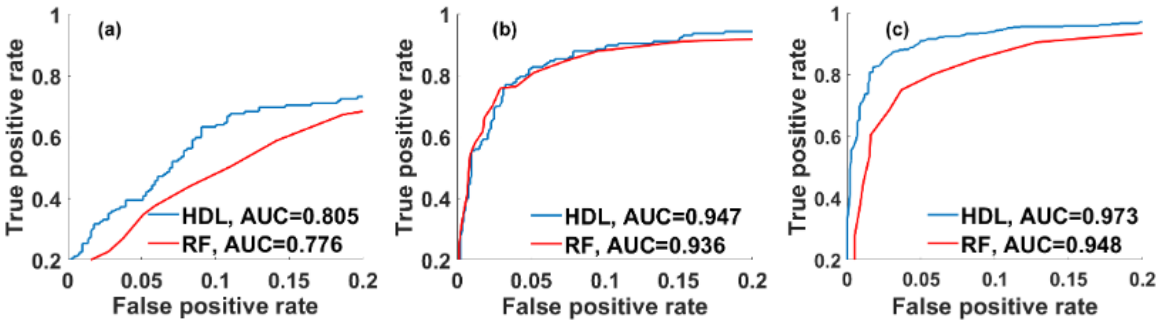

Figure 3-9: The ROC curves from test sets for HDL model and the Random Forest (RF) model, for C-Mod (a), DIII-D (b) and EAST (c). We only show the upper left region of the curves where the predictors have highest performance.

3.4.2 HDL performances on the three devices and benchmark with Random Forest

The HDL predictor successfully achieves state-of-the-art performance on all three test sets when compared to other fully-optimized deep NN disruption predictors [30]. To see this, we trained three HDL predictors (with fixed hyperparameters, as given in Section 3.4.1) and three Random Forest (RF) predictors [1, 4, 35] using the training set of each machine, and evaluated their performances on the test set corresponding to that machine. These results are shown in Figure 3-9. To carry out a fair comparison with previous approaches, the RF predictors for each machine are specifically optimized using the corresponding validation set: we carried out a K-fold cross validation procedure together with a parallelized grid search to find the optimal set of time threshold and forest hyperparameters for each machine using a binary classification performance metric called the F_γ-score [1, 4]. The HDL predictor exceeds RF performance on all three datasets: it triggers fewer false alarms on good discharges while missing fewer real disruptions. this shows the strong applicability and generalization power of the model. This general improvement on multiple machines seems mainly to come from the advantage of the sequence-based model that is designed for time series processing, as suggested in Section 3.3. Besides its impressive performance, the inference time of our model is short, allowing it to make a prediction in roughly 1ms using an 8-core CPU. The development of this fast and novel model constitutes an important step toward the prediction requirements of future devices. It also suggests a powerful conclusion: a common set of model hyperparameters used for three predictors can achieve high performance on all three machines, suggesting that although different devices may have disjoint operational regimes, there seems to exist a common type of discriminant function – with the same model hyperparameters - capable of separating the disruptive from non-disruptive phases on all these machines.

3.5 HDL cross-machine study on Alcator C-Mod, DIII-D, and EAST data

The availability of extensive experimental data sets covering several tokamaks allows us to design numerical experiments to test the transfer learning capabilities of the HDL architecture. Future reactors like ITER cannot tolerate more than a few unmitigated disruptions [36], so we must be able to predict their disruptions given very limited disruptive data from the reactors themselves. Expanding from the cross-machine DL-based disruption prediction study, we have designed complete numerical experiments to test the transfer learning capabilities of the HDL architecture. Given the availability of a large database of aggregated data from very different tokamaks, it is important to verify if, and how usefully the data from existing devices can be used to predict unstable plasmas on a new device. In this section, we consider two machines as *'existing machines'* and investigate the effect of their data for the HDL disruption predictor when used on the third machine, chosen as a *'new device'*. We primarily focus on the EAST case (EAST is chosen as the *'new device'*) in the following section. However, **all of the resulting qualitative conclusions are machine-independent**. They always hold, no matter which device is selected as the *'new device'*. Results regarding the other two case permutations can be found in appendix A.6 of [22].

3.5.1 Cross-machine prediction performance using the HDL architecture

As a first step, we would like to test the cross-machine performance of the HDL model. To do this, we train the HDL network using data from two *'existing devices'*, and test its performance on the third, unseen, *'new device'*. Following the predictors "with a glimpse" or "from scratch" approaches [30, 37], we then add 10 disruptive and 10 non-disruptive discharges from the *'new device'* to the training sets and do indeed observe a boost in the test set performances when using limited data from the target device. In the context of previous cross-machine studies [30, 37], our HDL framework shows promising transferrability on these three different devices; these test results can be found in Table 3.3 (all values reported here are AUCs averaged over the network ensemble).

Beyond performance, we are interested in investigating how data from different existing devices influence predictions of disruptions on a new one, and in particular if any effect can be linked to general, device-independent knowledge. To this aim, we design two further sets of cross-machine numerical experiments. The training set composition for each experiment can be found in Table 3.4.

Table 3.3: Cross machine prediction results of HDL

Training set	C-Mod +DIII-D	C-Mod +DIII-D + few EAST data	EAST +C-Mod	EAST +C-Mod + few DIII-D data	EAST +DIII-D	EAST +DIII-D + few C-Mod data
Test set	EAST	EAST	DIII-D	DIII-D	C-Mod	C-Mod
HDL Ensemble	0.788	0.819	0.622	0.741	0.564	0.605
HDL Ensemble + CMLS	0.806	0.837	0.659	0.765	0.588	0.631

Table 3.4: Training set composition of all cross machine experiments using EAST as the *'new machine'*

	Existing machines (C-Mod+DIII-D)		New machine (EAST)	
Case NO.	Non-disruptive	Disruptive	Non-disruptive	Disruptive
1	None	All (692+732)	All (5995)	20
2	None	All	All	None
3	None	All	50% (2998)	20
4	None	None	All	20
5	All (2651+4554)	All	All	20
6	All	All	None	None
7	None	All	All	All (2301)
8	All	All	All	All
9	None	None	All	All
10	None	All	$\approx$ 33% (1998)	All
11	$\approx$ 20% (692+732)	None	$\approx$ 33%	All
12	None	None	$\approx$ 33%	All

Values in parentheses give the exact number of shots.

3.5.2 Cross-machine experiments using limited disruptive data from the *'new device'*

The first set of cross-machine experiments was conducted using limited disruptive training shots from the new device. The results of these numerical experiments are shown in Figure 3-10(a)-(b). In the first experiment, the disruption predictor is trained on 20 randomly selected disruptive training shots and all non-disruptive training shots from the target new device, plus disruptive shots from two other devices (existing machines). This combination achieves the best performance on the new device test datasets (AUC=0.959, for the EAST case). In the second and third experiment, we first remove all new device disruptive shots and then 50% of new device non-disruptive shots from the first training dataset, separately. In the fourth experiment, the predictor was trained only using selected new device training data (20 disruptive training shots, all non-disruptive training shots), this being our limited data baseline model. In the fifth experiment, we add non-disruptive shots from two other machines to the first training dataset. In the sixth experiment, the predictor is trained only on data from 'existing' machines (no new device data) and its low performance highlights the importance of adding non-disruptive data from the 'new' target machine. From these numerical experiments, it is possible to draw the following conclusions:

- HDL achieves relatively good performances for a new device by including a few disruptive shots and many non-disruptive shots from the new device, plus many disruptive shots from existing devices. All components mentioned above are necessary because removing any of them significantly decreases the performance (cases 1 to 4 in Figure 3-10(a)).
- Non-disruptive data from existing devices is harmful to HDL performance, while disruptive data from existing devices improves the predictive power (cases 1, 4, 5 in Figure 3-10(b)).
- Non-disruptive data from the 'new' target device can substantially improve the predictive power (case 6 in Figure 3-10(b)).

3.5.3 Cross-machine experiments using all disruptive data from the *'new device'*

To further investigate the effect of the class imbalance in the training set, we conducted another set of experiments using all disruptive training shots of the new device. The results are reported in Figure 3-11. In this seventh experiment, the disruption

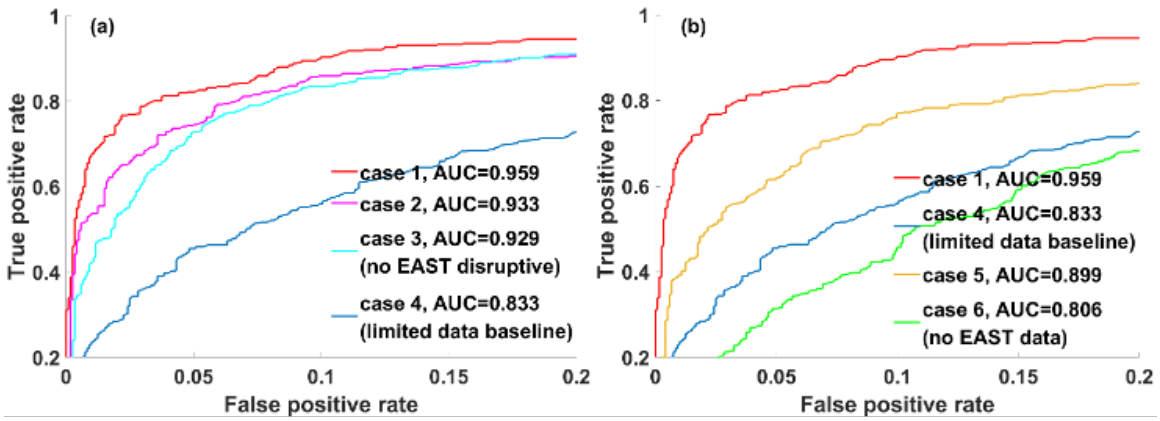

Figure 3-10: ROC curves from the EAST test using limited EAST disruptive training data.

predictor is trained on all disruptive and non-disruptive training shots from the new device, including disruptive shots from two other machines, and it achieves the best performance on the new device test dataset (AUC=0.983, for the EAST case). In the eighth experiment, we add non-disruptive shots from two other machines to the first training dataset. In the ninth experiment, the predictor is trained only on all new device training data, comprising the all-data baseline case for comparison. In experiments 10-12 (Figure 3-11(b)), we randomly remove most of the new device non-disruptive training shots, thus reducing the new device non-disruptive training data to be less than new device disruptive training data, i.e. an inversely imbalanced situation. The test results from Figure 3-11(a)-(b) point to the following further conclusions:

- Adding disruptive data from existing machines can still slightly improve test performances on the new device even if you have abundant new machine data (cases 7, 9 in Figure 3-11(a)). However, adding non-disruptive data from existing machines is still harmful in this situation (cases 7, 8 in Figure 3-11(a)).

- The effects of disruptive data (positive) and non-disruptive data (negative) do not result from the class imbalance of the new machine dataset, because disruptive data from existing devices continually have positive effects, while adding existing device non-disruptive data still has negative effects in the inversely imbalanced situation (Figure 3-11(b)). This difference between disruptive and non-disruptive data is machine independent, i.e. a universal conclusion.

- Removing non-disruptive data from the target device always decreases the test performance, no matter how imbalanced the target dataset is (cases 1, 3 in Figure 3-10(a), case 9, 12 in Figure 3-11(b)).

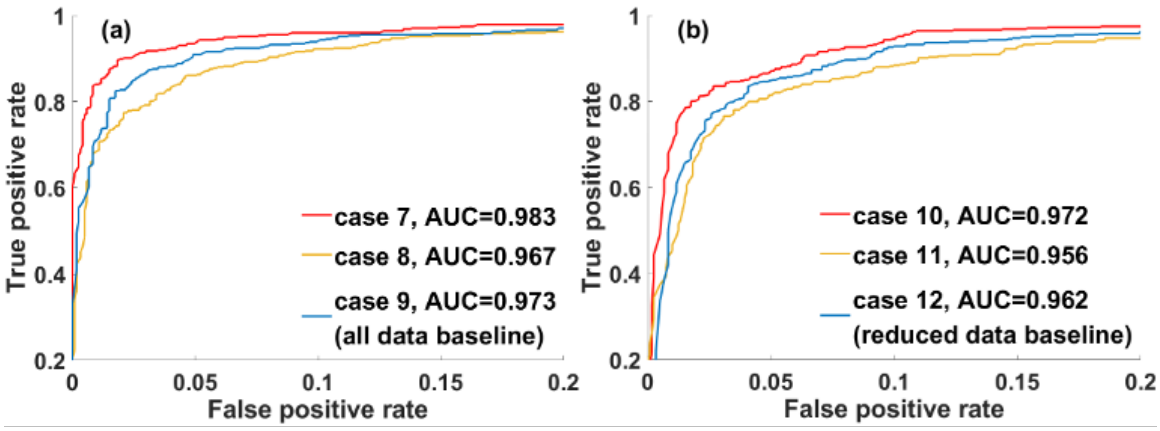

Figure 3-11: ROC curves from the EAST test set using all EAST disruptive training data.

3.5.4 Summary of Conclusions for Cross-machine Numerical Experiments

Considering all the conclusions in section Section 3.5.2 and Section 3.5.3, it is possible to state that knowledge of disruptive data from existing devices improves the performance on the new device while the non-disruptive data seem to have negative effects, which do not result from the label imbalance of training datasets. This suggests that the non-disruptive data are specific to one device, but disruptive data contain some general knowledge about disruptions dynamics that could be transferred to a new device, when using predictive, data-driven models. Indeed, different machines usually have different operational spaces, spatiotemporal scales for physics events and plasma diagnostics [1, 4, 30]. In other words, the distributions of plasma signals can vary significantly from one machine to others. From the data-driven perspective, this further implies that finding a numerical transformation that maps a set of signals from one device to a different device can be very challenging without incorporating machine-specific information, and this might indeed pose a great challenge when comparing ITER's operational space to all existing devices. Due to these considerations, we conclude that non-disruptive data from existing devices are machine-specific and will only decrease the accuracy of the predictive models on the new device when they are directly mixed with data from the target device. Nevertheless, different devices show similar behavior when operating close to a disruption. For example, the plasma internal inductance [1], the loop voltage and the locked mode signals [1, 4] have been observed to consistently increase on multiple machines when disruptions are imminent. These universal trends can be well captured by our time sequence based model as general knowledge about disruptions hidden beneath the disruptive data, and they help disruption prediction on new devices.

3.6 Summary

In this chapter, we have discussed findings from an explorative data analysis study on a C-Mod disruption database using a dimensionality-reduction technique called t-SNE, and demonstrated that time sequence data can better separate the disruptive and non-disruptive behavior compared to the instantaneous plasma state data (i.e. individual time slices). Based on these findings, we have designed a new, more powerful disruption prediction algorithm based on Deep Learning and also demonstrated a general, effective way to transfer knowledge from existing devices to new devices which offers guidelines for disruption prediction for a new device using onlt limited disruptive data from that device. The cross-machine study on Alcator C-Mod, DIII-D, and EAST shows that, given the highly elaborated deep learning architecture, it is not enough to use only data from existing devices to predict disruptions on a new tokamak device. The numerical experiments discussed in Section 3.5 demonstrate that, when compared with models using only data from existing devices, the model's performance greatly improves if both non-disruptive and some disruptive data are included from the target device. In particular, the HDL predictors can reach AUC $>$ 0.95 on EAST if trained including only a small set (20) of disruptive discharges from the target device (EAST), while simultaneously using all available non-disruptive information from the target machine.

Furthermore, disruptive and non-disruptive data are found to have different impacts on the cross-machine disruption prediction framework, with the implication that non-disruptive data are machine-specific while disruptive data contain more general knowledge about disruptions. These results are an important milestone for disruption prediction research for application to next-generation burning plasma reactors, such as ITER.

References - Chapter 3

[1] C. Rea, R. S. Granetz, K. Montes, R. A. Tinguely, N. Eidietis, J. M. Hanson, and B. Sammuli. Disruption prediction investigations using Machine Learning tools on DIII-D and Alcator C-Mod. *Plasma Physics and Controlled Fusion*, 60(8):084004, 8 2018.

[2] Cristina Rea and Robert S. Granetz. Exploratory Machine Learning Studies for Disruption Prediction Using Large Databases on DIII-D. *Fusion Science and Technology*, 74(1-2):89–100, 8 2018.

[3] C. Rea, K. J. Montes, A. Pau, R. S. Granetz, and O. Sauter. Progress toward interpretable machine learning-based disruption predictors across tokamaks. *Fusion Science and Technology*, 2020.

[4] K. J. Montes, C. Rea, R. S. Granetz, R. A. Tinguely, N. Eidietis, O. M. Meneghini, D. L. Chen, B. Shen, B. J. Xiao, K. Erickson, and M. D. Boyer. Machine learning for disruption warnings on Alcator C-Mod, DIII-D, and EAST. *Nuclear Fusion*, 59(9):096015, 2019.

[5] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[6] George V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

[7] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[9] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[10] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.

[11] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3):121–136, 1975.

[12] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[13] Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. 04 1991.

[14] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with

gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[16] Kyunghyun Cho, Bart Merrienboer, Dzmitry Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. 09 2014.

[17] Shudong Yang, Xueying Yu, and Ying Zhou. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, pages 98–101, 2020.

[18] S.P. Gerhardt, D.S. Darrow, R.E. Bell, B.P. LeBlanc, J.E. Menard, D. Mueller, A.L. Roquemore, S.A. Sabbagh, and H. Yuh. Detection of disruptions in the high- β spherical torus NSTX. *Nuclear Fusion*, 53(6):063021, 6 2013.

[19] Jesús Vega, Sebastián Dormido-Canto, Juan M. López, Andrea Murari, Jesús M. Ramírez, Raúl Moreno, Mariano Ruiz, Diogo Alves, and Robert Felton. Results of the JET real-time disruption predictor in the ITER-like wall campaigns. *Fusion Engineering and Design*, 88(6-8):1228–1231, Oct 2013.

[20] C. Rea, K. J. Montes, K. G. Erickson, R. S. Granetz, and R. A. Tinguely. A real-time machine learning-based disruption predictor in DIII-D. *Nuclear Fusion*, 59(9):096016, 2019.

[21] A. Pau, A. Fanni, S. Carcangiu, B. Cannas, G. Sias, A. Murari, and F. Rimini. A machine learning approach based on generative topographic mapping for disruption prevention and avoidance at JET. *Nuclear Fusion*, 59(10):106017, 2019.

[22] J. Zhu, C. Rea, K. J. Montes, R. Granetz, R. Sweeney, and R. A. Tinguely. Hybrid deep learning architecture for general disruption prediction across tokamaks. *Nuclear Fusion*, 61(2):026007, 2020.

[23] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[24] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. in eighth workshop on syntax. *Semantics and Structure in Statistical Translation (SSST-8), 2014*, 2014.

[25] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017.

[26] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[28] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[29] E. M. Hollmann, P. B. Aleynikov, T. Fülöp, D. A. Humphreys, V. A. Izzo, M. Lehnen, V. E. Lukash, G. Papp, G. Pautasso, F. Saint-Laurent, and J. A. Snipes. Status of research toward the ITER disruption mitigation system. *Physics of Plasmas*, 22(2):021802, 2 2015.

[30] Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568:526–531, 2019.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[32] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[33] Simon Haykin. Neural networks: A comprehensive foundation, 3rd edn. 1999.

[34] MP Perrone and LN Cooper. $^{\underline{a}}$when networks disagree: Ensemble methods for hybrid neural networks, $^{\underline{o}}$ neural networks for speech and image processing. *Chapman-Hall*, 1993.

[35] Yichen Fu, David Eldon, Keith Erickson, Kornee Kleijwegt, Leonard Lupin-Jimenez, Mark D. Boyer, Nick Eidietis, Nathaniel Barbour, Olivier Izacard, and Egemen Kolemen. Machine learning control for disruption and tearing mode avoidance. *Physics of Plasmas*, 27(2):022501, 2020.

[36] P. C. de Vries, G. Pautasso, D. Humphreys, M. Lehnen, S. Maruyama, J. A. Snipes, A. Vergara, and L. Zabeo. Requirements for triggering the ITER disruption mitigation system. *Fusion Science and Technology*, 69(2):471–484, 2016.

[37] G. A. Rattá, J. Vega, and A. Murari. Viability Assessment of a Cross-Tokamak AUG-JET Disruption Predictor. *Fusion Science and Technology*, 74(1-2):13–22, 8 2018.

Chapter 4

Scenario Adaptive Disruption Prediction

Next generation high performance (HP) tokamaks risk damage from unmitigated disruptions at high current and power. Achieving reliable disruption prediction for a device's HP operation based on its low performance (LP) data is key to success. In this chapter, through explorative data analysis and dedicated numerical experiments utilizing data from multiple existing tokamaks, we demonstrate how the operational regimes of tokamaks can affect the power of a trained disruption predictor. First, our results suggest data-driven disruption predictors trained on abundant LP discharge data work poorly for the HP regime of the same tokamak, which is a consequence of the distinct distributions of the tightly correlated signals related to disruptions in these two regimes. Second, we find that matching operational parameters among tokamaks strongly improves cross-machine accuracy, which implies our model learns from the underlying scalings of dimensionless physics parameters like q_{95} and β_p, and confirms the importance of these parameters in disruption physics and cross machine domain matching from the data-driven perspective. Finally, our results show how, in the absence of HP data from the target devices, the best predictivity of the HP regime for the target machine can be achieved by combining LP data from the target with HP data from other machines. These results provide a possible disruption predictor development strategy for next generation tokamaks, such as ITER and SPARC, highlighting the importance of exploring, on existing machines, baseline scenario discharges expected in future tokamaks. In this way, it should be possible to collect the relevant disruptive data which can be used to refine the disruption prediction models for the future devices.

4.1 Introduction and Motivation

In chapter 3, we demonstrated that data-driven disruption predictors, especially those using deep-learning based models, can achieve high accuracy on existing tokamaks. However, as discussed in Section 3.5, due to the large gap of dimensional and operational parameters between existing devices and next generation tokamaks, extrapolation of these predictors to near-future burning-plasma tokamaks, like ITER [1] and SPARC [2], is uncertain. For example, the HDL model trained on DIII-D and EAST data can only achieve AUC=0.588 for C-Mod. So far, significant effort has been devoted to solving this problem. First, recent deep-learning based predictors have shown strong cross-machine ability [3, 4] to learn general representations across tokamaks. Second, the strategy of building a predictor from scratch is proposed by several existed analysis [5, 6]. These studies explored the strategy of progressively retraining the predictor on historical data and then testing on future unseen discharges. However, according to our discussion in Section 2.3, the key assumption of these studies that the parameters of future unseen discharges are included in the operational regime of historical data might not hold for for future devices like ITER. During the performance ramp-up phase of these future devices, data-driven predictors trained on data from initial LP campaigns might not work well on subsequent HP campaigns, due to the shift of plasma parameters, and thus the predictors trained on LP discharges are likely to be ineffective for HP discharges. In this chapter, we will look into how the performance of the trained predictors change when the test scenario deviates from the scenarios included in the training set. Our goal is to find a training strategy for a data-driven predictor that works for the HP regime of the target device, while only using LP data from the target device, combined with selected data from other tokamaks. The numerical experiment described in this chapter is focused on ITER, but the final strategy should also be applicable to other future devices.

4.2 Using data from existing machines to simulate the LP and HP phases on ITER

As currently envisioned, the ITER research plan incorporates a staged approach strategy aiming to increase the experimental capabilities in phases leading up to HP fusion operation [1, 7]. Following the completion of its first plasma, ITER will increase the toroidal magnetic field (B_{tor}), plasma current (I_p), density, and input power (P_{in}) toward final high I_p, B_{tor} and P_{in} fusion operation. Ramping up these parameters during ITER's early operation can pose challenges to the development of disruption predictors. First, the distributions of operational parameters - such as normalized

plasma pressure (β_p), safety factor at the 95% flux surface (q_{95}), and Greenwald density (n_G) - will change with the increasing I_p, B_{tor} and P_{in} any or all of which might impact the efficacy of any trained predictor. Second, due to the potentially serious damage to the device from high-current, high-stored-energy disruptions, the ITER research plan requires developing a reliable Disruption Mitigation System (DMS) trigger before the beginning of HP operation [7]. This requirement can result in differences between the training and testing operational regimes of the DMS trigger and may invalidate the disruption prediction algorithm.

To simulate possible discrepancies between the training and testing domains of the ITER DMS trigger, we select three parameters: β_p, P_{in} (not a training feature) and q_{95} that are closely related to tokamak operation but less significant to disruption prediction for the three tokamaks we studied [4] and calculate their I_p-flattop-averaged values[1] for each plasma discharge in our databases. From the distributions of flattop averaged parameters, we choose low/high cutoff thresholds for each of the three parameters (Table 4.1) and select various LP/HP (low/high β_p, low/high P_{in}, high/low q_{95}) datasets on three devices based on the ranges of these three parameters. The chosen cutoff thresholds vary for different devices and depend on the distributions of each signal on the different devices as well as typical operational scenarios on these devices. Notice that the three chosen parameters are a small subset of all signals used for prediction models [8]. To see how limiting the ranges of the three chosen parameters affects the distributions of other training features, an orthogonal linear transformation, called Principal Component Analysis (PCA) [9], is applied to all 12 training features (including q_{95} and β_p) of the combined LP and HP datasets for all three devices. Before applying the PCA transformation, each signal in this combined dataset is separately normalized to mean= 0, and standard deviation= 1, such that two principal components are not dominated by q_{95} and β_p. In Figure 4-1, each magenta point represents a 10 time-step sequence of 12 training features, randomly sampled from the flattop of an HP shot, while each cyan point represents a sequence randomly sampled from the flattop of an LP shot. The two principal components (x, y axes) are linear combinations of 12 training features, and our PCA suggests that the 10 unconstrained features make significant contributions. If the joint distribution of unconstrained features is not strongly affected by three chosen parameters (similar for LP and HP plasmas), the distributions of resulting LP and HP plasmas in the projected 2-D plane should have a large overlap. However, the PCA clustering plots show that there is only a very small overlap between the resulting LP/HP plasmas for all three devices, implying thatsignals related to disruption prediction are closely correlated. Limiting the ranges of a few of the less significant parameters can signifi-

[1]For β_p and P_{in}, the average is only computed during the flattop I_p period when external heating is active

Table 4.1: Performance cutoff threshold of β_p, P_{in} and q_{95} on three devices

Device	β_p low/high cutoff		P_{in} low/high cutoff (MW)		q_{95} low/high cutoff	
C-Mod	<0.15	>0.25	<1.0	>3.0	<4.0	>4.6
DIII-D	<0.60	>0.80	<3.5	>7.5	<4.5	>5.0
EAST	<0.55	>0.75	<0.6	>3.0	<5.0	>6.0

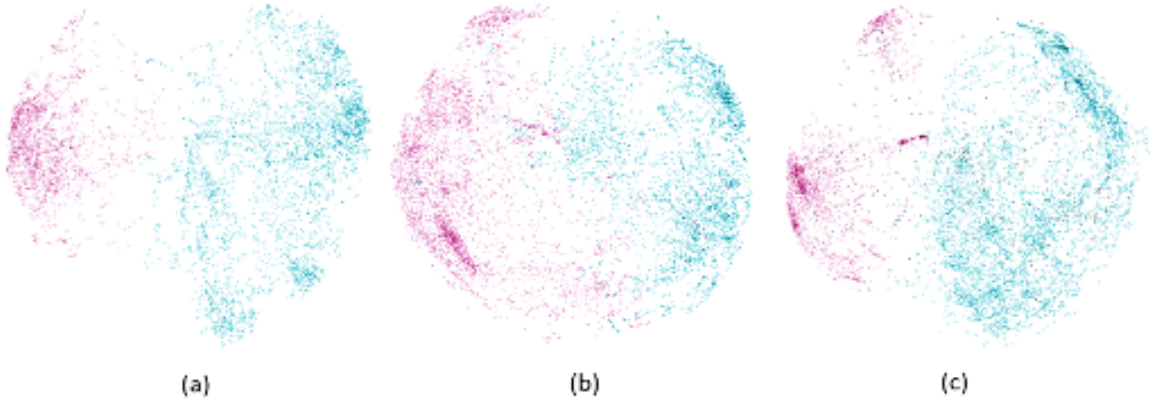

Figure 4-1: The PCA clustering plots for: (a) C-Mod; (b) DIII-D; and (c) EAST. Each magenta point represents a 10 time-step sequence of 12 training features randomly sampled from the flattop of a HP shot while each cyan point represents a sequence randomly sampled from the flattop of a LP shot. The coloring is done *a posteriori*.

cantly change the distributions of other signals related to disruption prediction, and makes clear the distinction between LP and HP plasmas. This observation further implies that the LP regime physics is too limited and does not have enough overlap with the HP regime physics to adequately train the predictor. More PCA plots, for different subdivisions of HP data, can be found in [10] which further support our conclusion about the signal correlation.

Another component of our study is the disruption predictor. In our previous research, we have developed a Hybrid Deep-Learning (HDL) disruption predictor that achieves state-of-the-art accuracy on multiple tokamaks, with only limited hyper-parameter tuning [8]. Throughout this chapter, we will use the HDL predictor to conduct all numerical disruption studies, and the result of each experiment is evaluated using the ROC curves at 50ms before the final current quench. However, since all data-driven methods essentially learn from the empirical distribution of the input signals, we argue that our analysis is generally applicable to all data-driven methods. 50 ms is chosen since it is the warning time required for the ITER DMS [11].

4.3 Scenario based cross-machine study

Given the fact that unmitigated high-current, high-stored-energy disruptions can seriously threaten the integrity of future burning plasma tokamaks, developing a disruption predictor for HP operations burning-plasma of these devices with only LP data from themselves, is one of the suggested approaches for future burning plasma devices. Based on this approach, a *"train-on-LP-data"* strategy, as described in [5, 6], consists of training a predictor using data from the early stages of ITER's operation, and then applying it to subsequent discharges. If a predictor trained on initial LP ITER data has sufficient knowledge that is applicable to the HP regime, it should be able to predict disruptions in the HP regime. Using an HDL predictor, and various LP/HP datasets from three existing tokamaks, we investigate whether this *"train-on-LP-data"* strategy works. If not, given the strong cross-machine potential of deep-learning-based predictors, we seek to improve target prediction accuracy by using data from other devices. Here, we consider C-Mod and EAST as *'existing/other machines'*, with DIII-D chosen as the *'new/target device'* and conduct numerical experiments to explore the best strategy of developing data-driven predictors that can predict disruptions in the HP regime of the *new device*, using only LP data from the *new device*, combined with HP data from the *existing machines*. The training and testing set compositions of all experiments can be found in Table 4.2[2] [3]. In addition, **all following qualitative conclusions are machine-independent**: they always hold no matter which device is selected as the *'new device'*. The other two permutations are shown in [10].

The first set of numerical experiments is conducted using only data from our target *new device* to test the effectiveness of the *"train-on-LP-data"* strategy. The results of these experiments (cases 1-5) are shown in Figure 4-2(a)-(b). The training and testing set composition of these cases can be found in Table 4.2. From the results of these cases, it is possible to draw the following conclusions:

- Limiting the ranges of chosen parameters in the training set strongly affects the test performance of a trained data-driven predictor. A predictor trained on a few hundred high q_{95} discharges works poorly for the HP regime of the same device (case 2 in Figure 4-2(a)). Furthermore, as more parameters (β_p, P_{in}) of the training discharges deviate from the target HP regime, the predictor's

[2]For case 3, 140 DIII-D HP shots were selected from the total 240 DIII-D HP shots as the test set, and the remaining 100 DIII-D HP shots were used for the training set. 11 independent experiments were run for case 3 (11 different random partitions of 240 DIII-D HP shots). The case 3 result, shown in Figure 3-6, corresponds to the median accuracy among 11 results, making it comparable with the results for the other cases.

[3]For cases 4-5, 140 DIII-D HP high B_{tor} shots were selected from 240 DIII-D HP shots as the test set to evaluate the effect of B_{tor}.

Table 4.2: Training and testing set composition of all experiments using DIII-D as the *'new machine'*

Case No.	Training set	Test set
1	209 DIII-D LP (β_p <0.6, P_{in} <3.5MW, q_{95} >5) shots (14% disruptive)	240 DIII-D HP (β_p >0.8, P_{in} >7.5MW, q_{95} <4.5) shots (15% disruptive)
2	209 DIII-D high q_{95} (q_{95} >5) shots (14% disruptive)	
3	100 DIII-D HP shots (15% disruptive)	140 DIII-D HP shots (15% disruptive)
4	100 DIII-D HP, low B_{tor} (<1.79T) shots (20% disruptive)	140 DIII-D HP, high B_{tor} (>1.79T) shots (12% disruptive)
5	100 DIII-D high β_p (>0.8), low q_{95} (<4.5), low B_{tor} (<1.79T) shots (14% disruptive)	
6	209 C-Mod low q_{95} (<4) shots	Same as cases 1-2
7	209 C-Mod high q_{95} (>5) shots	
8	209 EAST low q_{95} (<5) shots	
9	209 EAST high q_{95} (>6) shots	
10	209 EAST low q_{95} (<5), high β_p (>0.4) shots	
11	209 C-Mod low q_{95} (<4) shots plus 209 DIII-D LP shots	
12	209 EAST low q_{95} (<5), high β_p (>0.4) shots plus 209 DIII-D LP shots	
13	209 C-Mod high q_{95} (>5) shots plus 209 DIII-D LP shots	
14	209 EAST high q_{95} (>6) shots plus 209 DIII-D LP shots	

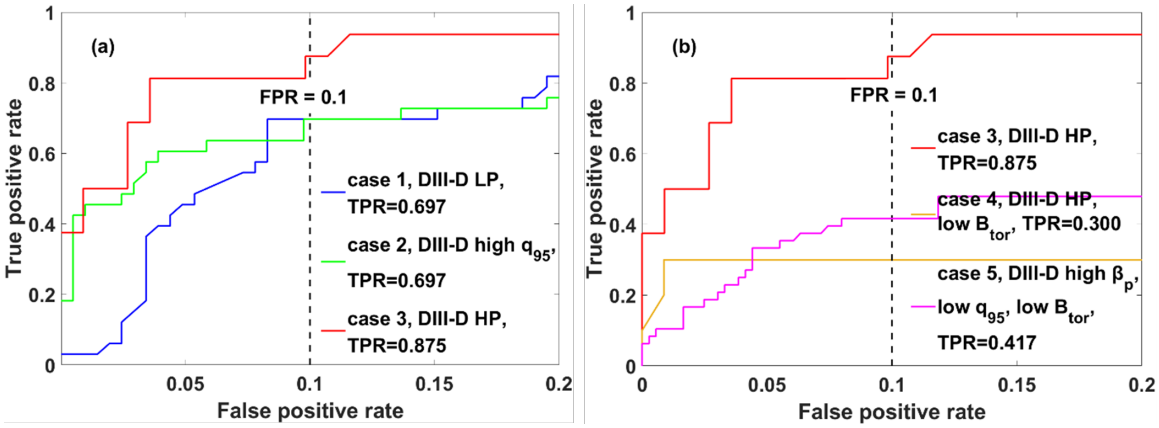

Figure 4-2: ROC curves from the *new device* (DIII-D) test set using only *new device* data. The training and testing set compositions of all cases can be found in Table 4.2.

accuracy for the HP regime becomes systematically worse (case 1 in Figure 4-2(a)). Notice that there are different numbers of disruptive discharges in the different training sets. Although having only a small number of disruptive samples in the training set can decrease the accuracy of a trained predictor, our results suggest this is a secondary effect compared with the effects of changing the operational regime. Despite having the most disruptive training shots (30) in case 1, this case gives the worst test accuracy among cases 1-3. Given these results, we conclude that a predictor trained only on abundant LP discharges performs poorly for the HP regime of the same device.

- A data-driven predictor can effectively learn disruption physics if the training and test data come from similar operational regimes. A predictor trained on only 100 HP shots of the target device already achieves the best test accuracy among cases 1-3 (case 3 in Figure 4-2(a)-(b)).

From the first conclusion above, even without other constraints to the training set, the q_{95} discrepancy between the training and testing sets can significantly decrease the prediction accuracy for the target HP regime. Since high current disruptions can be dangerous to ITER, we want to develop a predictor using only low current ITER discharges. Under this constraint, to match q_{95} between training and testing regimes, one approach is to train a predictor on low B_{tor}, low current, and thus low q_{95}, discharges. To test this, we sub-select low B_{tor} shots from the *new device* HP database as the training set, and test on the remaining HP high B_{tor} *new device* data (case 4). However, selecting high P_{in} shots from low B_{tor} discharges can yield highly skewed dataset[4], in the fifth case, the predictor is trained on low q_{95}, high β_p and low

[4]for example, C-Mod can not run shots with ICRF heating when the B_{tor} is low

B_{tor} *new device* shots, and tested on HP high B_{tor} *new device* data. In Figure 4-2(b), we compare the results of cases 4 and 5 with case 3 (training and testing data from the same HP regime) which yields the following additional conclusions:

- Although P_{in} and B_{tor} are not training features, predictors trained on HP (with/without P_{in} constraint) low B_{tor} discharges perform poorly for the HP high B_{tor} discharges. This implies that the ranges of parameters like B_{tor} and P_{in} can greatly affect the feature space of predictors, even when they are not training features. This indicates a need for ITER to reach relatively **high** B_{tor} as early as possible during its LP pre-fusion phase, even with low current and high q_{95} (cases 3-5 in Figure 4-2(b)).

From the existing literature [4] and our previous HDL studies [8], q_{95} and β_p are not the signals with the most significance for disruption prediction on the three tokamaks we studied. Therefore, directly learning from these constrained features is not required for achieving high prediction rate and the discrepancies of P_{in}, q_{95} and β_p ranges between LP and HP data themselves will not lead to significantly worse prediction accuracy on the test set. Since most of the training features (10/12) are not artificially constrained, if limiting the ranges of three chosen parameters does not significantly change distributions of other parameters, the predictor trained on the resulting LP data (case 1) should work well on HP test set (close to the result of case 3). However, the above results show that predictor trained on LP data works significantly more poorly on an HP test set. This observation again suggests signals related to disruption prediction are strongly correlated. Although the chosen physics-based signals (β_p, P_{in}, q_{95}) do not directly contribute significantly to the power of the model, limiting their ranges can strongly affect the distributions of more important signals and hence change the prediction results. Thus, without additional data, developing a disruption predictor that works well for the HP regime of a given tokamak, using only LP data from that same device, is unlikely to succeed because the LP regime physics is too limited to yield good predictions for disruptions in the HP regime.

To seek a better strategy, we conducted another set of numerical experiments, using data from both the new device and existing machines. Given the results from the first set of numerical experiments, an attempt was made to match parameters between the new device and the existing machines. The results of these experiments (cases 6-12) are shown in Figure 4-3(a)-(d). The training and testing set composition of these cases can be found in Table 4.2. The results of figure 3 point to the following conclusions:

- A predictor trained on low q_{95} data from existing machines performs better than a predictor trained on high q_{95} data from existing machines for the HP regime of

the new device (cases 6, 7 in Figure 4-3(a), cases 8, 9 in Figure 4-3(b)). Training a predictor using low q_{95} and relatively high β_p data from existing machines further increases the accuracy for the HP regime of the new device (case 10 in Figure 4-3(b)). These results demonstrate that training on "matched" data (with similar operational parameters to those in the test dataset) from existing machines greatly outperforms the unmatched data, and progressively matching more operational parameters continuously improves the target performance. Therefore, developing ITER baseline scenario discharges on existing tokamaks, and training predictors on these, should greatly improve disruption prediction for ITER itself.

- In the absence of HP data from the new device, combining "matched" HP data from existing machines with LP data from the new device gives the best prediction rate for the HP regime of the new device, while adding "unmatched" data from existing machines to the training set can even decrease prediction accuracy on the target device (case 11, 13 in Figure 4-3(c), case 12, 14 in Figure 4-3(d)).

Considering the results from these two sets of numerical experiments, we conclude that, due to the distinct distributions of the tightly correlated signals related to disruptions in HP and LP regimes, any data-driven predictors trained on early LP ITER data cannot be directly applied to future HP operation. Our analysis shows that developing a reliable DMS trigger for ITER's HP operation, using only LP data from ITER itself, requires the addition of HP ITER baseline discharges from existing machines. A possible strategy for ITER DMS trigger development is as follows: combine ITER LP data (low β_p, low P_{in}, high q_{95} with relatively high B_{tor}) with HP ITER baseline discharges from other devices to train a predictor with enough accuracy to help ITER conserve its disruption budget during the early stage of its HP operation; as ITER's HP operation proceeds, add HP ITER data to the training set. Retraining the predictor using these combined datasets should boost the predictor's performance towards ITER's long-term requirements [12].

4.4 Summary and future plans

Given the risks of significant damage to fusion devices from unmitigated high-current, high-power disruptions, developing a DMS trigger for HP burning-plasma operation before starting HP campaigns is crucial for the success of next generation tokamaks. In this chapter, using databases from C-Mod, DIII-D and EAST, we selected three parameters that are closely related to tokamak operation, but less significant to disruption prediction on the three tokamaks and built LP/HP datasets that can simulate the LP and HP phases on ITER. Our preliminary data exploration using these

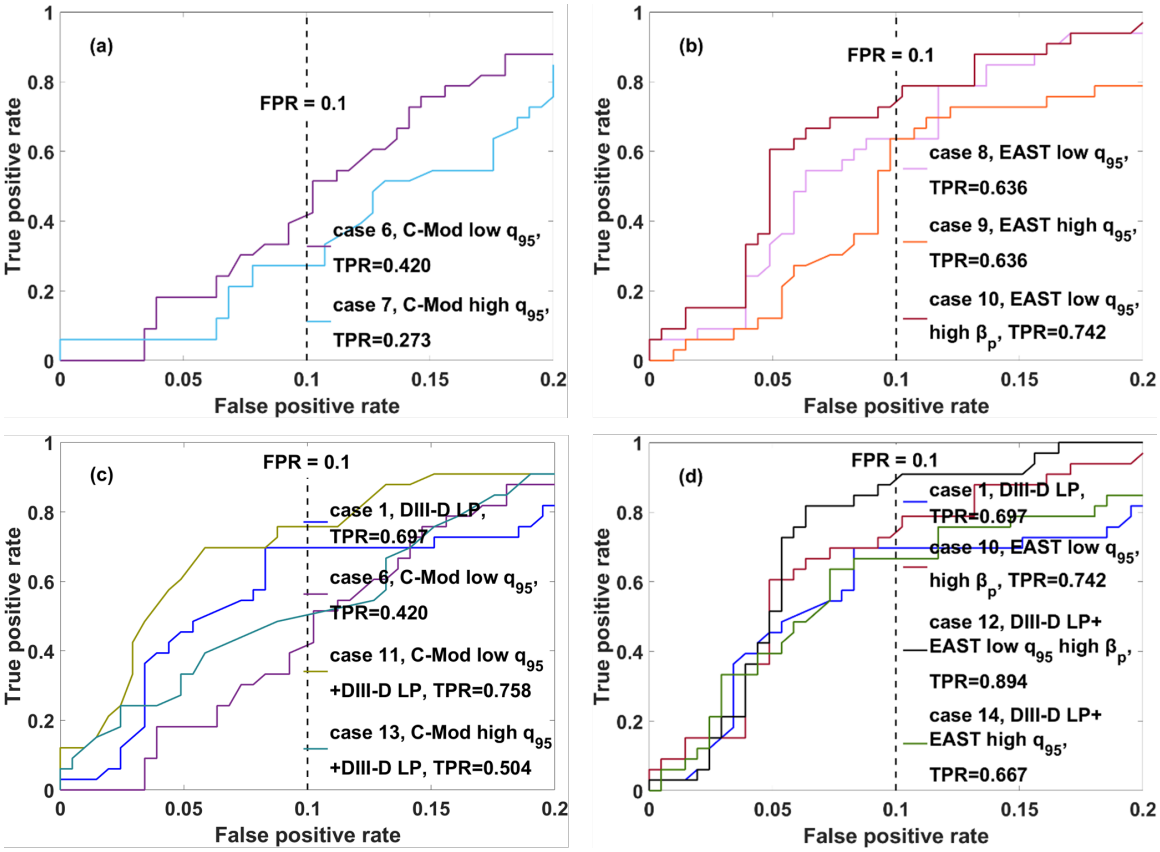

Figure 4-3: ROC curves from the *new device* (DIII-D) test set using both *new device* data and *existing machines* (C-Mod, EAST) data. The training and testing set compositions of all cases can be found in Table 4.2.

datasets finds that limiting the ranges of three chosen parameters clearly separates the resulting LP/HP plasmas regimes, and we find that using LP regime physics alone is insufficient for predicting HP regime disruptions. Dedicated numerical experiments based on these datasets further demonstrate that although a data-driven predictor can effectively learn when training and testing data come from the HP regime of the same device, having even one parameter of the training set deviate from the test operational regime greatly decreases the test performance of the trained predictors. Since q_{95} and β_p are not the most significant signals in the HDL model for the three machines we studied, the above results suggest that different signals related to disruption prediction are strongly correlated. Therefore, pushing the limits of less important signals changes the distributions of more significant signals and thus decreases the power of a trained predictor. Any data-driven predictors trained only on LP discharges perform poorly for the subsequent HP regime of the same tokamak, which suggests the *"train-on-LP-data alone"* strategy will not be sufficient for ITER.

Our cross-machine numerical experiments show that matching operational parameters among devices can greatly improve prediction accuracy for the target device. In the absence of HP data from the target device, the best prediction results for the HP regime of the target device can be achieved by training the predictor on LP data from the target combined with HP data from other machines. This conclusion implies that our model learns from the underlying scalings of dimensionless physics parameters, like q_{95}, and β_p and confirms the importance of these parameters in disruption physics and cross machine domain matching from the data-driven perspective. Given all above findings, we conclude that combining burning-plasma simulation discharges from experiments on existing tokamaks with initial LP data from the next step device is a promising strategy for the development of a DMS trigger for next step tokamaks. Thus, the development of a DMS trigger for future burning-plasma devices requires us to build comprehensive databases that consist of different kinds of disruptive burning-plasma baseline scenario discharges from current devices. Developing burning-plasma baseline scenarios on existing machines and exploring different kinds of disruptions that can happen during the next step device's HP operation, in the burning-plasma baseline scenarios of current devices, to collect relevant data, is crucial for improving disruption prediction on future devices.

References - Chapter 4

[1] T.C. Hender, J.C. Wesley, J. Bialek, A. Bondeson, A.H. Boozer, R.J. Buttery, A. Garofalo, T.P. Goodman, R.S. Granetz, Y. Gribov, O. Gruber, M. Gryaznevich, G. Giruzzi, S. Günter, N. Hayashi, P. Helander, C.C. Hegna, D.F. Howell, D.A. Humphreys, G.T.A. Huysmans, A.W. Hyatt, A. Isayama, S.C. Jardin, Y. Kawano, A. Kellman, C. Kessel, H.R. Koslowski, R.J. la Haye, E. Lazzaro, Y.Q. Liu, V. Lukash, J. Manickam, S. Medvedev, V. Mertens, S.V. Mirnov, Y. Nakamura, G. Navratil, M. Okabayashi, T. Ozeki, R. Paccagnella, G. Pautasso, F. Porcelli, V.D. Pustovitov, V. Riccardo, M. Sato, O. Sauter, M.J. Schaffer, M. Shimada, P. Sonato, E. J. Strait, M. Sugihara, M. Takechi, A.D. Turnbull, E. Westerhof, D.G. Whyte, R. Yoshino, and H. Zohm. Chapter 3: Mhd stability, operational limits and disruptions. *Nuclear Fusion*, 47(6):S128–S202, June 2007.

[2] A. J. Creely and M. J. et al. Greenwald. Overview of the SPARC tokamak. *Journal of Plasma Physics*, 86(5):865860502, 2020.

[3] Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568:526–531, 2019.

[4] K. J. Montes, C. Rea, R. S. Granetz, R. A. Tinguely, N. Eidietis, O. M. Meneghini, D. L. Chen, B. Shen, B. J. Xiao, K. Erickson, and M. D. Boyer. Machine learning for disruption warnings on Alcator C-Mod, DIII-D, and EAST. *Nuclear Fusion*, 59(9):096015, 2019.

[5] S. Dormido-Canto, J. Vega, J.M. Ramírez, A. Murari, R. Moreno, J.M. López, and A. Pereira. Development of an efficient real-time disruption predictor from scratch on JET and implications for ITER. *Nuclear Fusion*, 53(11):113001, 11 2013.

[6] J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, A. Pereira, and A. Acero. Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks. *Nuclear Fusion*, 54(12):123001, Dec 2014.

[7] DJ Campbell et al. The iter research plan. In *Proceedings of the 24th International Conference on Fusion Energy, San Diego, CA, USA*, pages 8–13, 2012.

[8] J. Zhu, C. Rea, K. J. Montes, R. Granetz, R. Sweeney, and R. A. Tinguely. Hybrid deep learning architecture for general disruption prediction across tokamaks. *Nuclear Fusion*, 61(2):026007, 2020.

[9] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[10] Jinxiang Zhu, Cristina Rea, RS Granetz, ES Marmar, KJ Montes, Ryan Sweeney, RA Tinguely, DL Chen, Biao Shen, BJ Xiao, et al. Scenario adaptive disruption prediction study for next generation burning-plasma tokamaks. *Nuclear Fusion*, 61(11):114005, 2021.

[11] E. M. Hollmann, P. B. Aleynikov, T. Fülöp, D. A. Humphreys, V. A. Izzo, M. Lehnen, V. E. Lukash, G. Papp, G. Pautasso, F. Saint-Laurent, and J. A. Snipes. Status of research toward the ITER disruption mitigation system.

Physics of Plasmas, 22(2):021802, 2 2015.

[12] P. C. de Vries, G. Pautasso, D. Humphreys, M. Lehnen, S. Maruyama, J. A. Snipes, A. Vergara, and L. Zabeo. Requirements for triggering the ITER disruption mitigation system. *Fusion Science and Technology*, 69(2):471–484, 2016.

Chapter 5

Integrated deep learning framework for unstable event identification and disruption prediction of tokamak plasmas

The ability to identify underlying disruption precursors is key to disruption avoidance. In this paper, we present an integrated deep learning (DL) based model that combines disruption prediction with the identification of several disruption precursors like rotating modes, locked modes, H-to-L back transitions and radiative collapses. The first part of our study demonstrates that the DL-based unstable event identifier trained on 160 manually labeled DIII-D shots can achieve, on average, 84% event identification rate of various frequent unstable events (like H-L back transition, locked mode, radiative collapse, rotating MHD mode, large sawtooth crash), and the trained identifier can be adapted to label unseen discharges, thus expanding the original manually labeled database. Based on these results, the integrated DL-based framework is developed using a combined database of manually labeled and automatically labeled DIII-D data, and it shows state-of-the-art (AUC=0.940) disruption prediction and event identification abilities on DIII-D. Through cross-machine numerical disruption prediction studies using this new integrated model and leveraging the C-Mod, DIII-D, and EAST disruption warning databases, we demonstrate the improved cross-machine disruption prediction ability and extended warning time of the new model compared with a baseline predictor. In addition, the trained integrated model shows qualitatively good cross-machine event identification ability. Given a labeled dataset, the strategy presented in this paper, i.e. one that combines a disruption predictor with an event identifier module, can be applied to upgrade any neural network based disruption predictor. The results presented here inform possible development strategies

of machine learning based disruption avoidance algorithms for future tokamaks and highlight the importance of building comprehensive databases with unstable event information on current machines.

As an introduction, Section 5.1 gives a motivation for the integrated DL model. The dataset used in the development and test of the integrated DL model is then described in Section 5.2. Following this, a detailed explanation of the iterative labelling process for assigning event labels to non-disruptive shots is presented in Section 5.3. Based on a fully labelled dataset, the development of an integrated DL model is presented in Section 5.4. Next, the results of cross-machine numerical experiments using the integrated model are discussed in Section 5.5. Finally, in Section 5.6, we present a discussion of conclusions and future plans with respect to the integrated DL model.

5.1 Introduction

As shown in chapter 2 and chapter 3, many disruption prediction studies [1–10] have proven the effectiveness of data-driven prediction methods. Furthermore, recent modeling efforts based on deep learning (DL) algorithms [9, 10] have shown improved performance and the potential cross-machine transferability of such predictive methods. However, DL approaches often lack the ability to identify disruption precursors, thus making them less explainable. This not only undermines the confidence of tokamak opera-tors in the results themselves but also hinders the implementation of disruption avoidance strategies.

On the other hand, according to our discussion in Section 2.1, there exist outstanding examples of physics-driven approaches to predict disruptions and their precursors [11–17]: the disruption event characterization and forecasting (DECAF) suite [15] incorporates various physics-based modules the identification and forecast of tearing modes, locked modes, resistive wall modes, edge localized modes (ELMs), among other unstable events. These physics modules are designed for stability boundary detection on different devices and some modules are accelerated via ML surrogate models [18]. Building upon these physics models, the DECAF suite can provide the proximity of plasma state to different disruption precursors and final disruption which gives it much better interpretability over data-driven methods and enables the machine operators to avoid disruptions instead of simply mitigating them.

In response to the need for unstable event identification via data-driven models, a integrated model that can detect several unstable events, and at the same time predict plasma disruptions, is developed using a manually labelled DIII-D dataset. Although both our integrated framework and physics-driven models like DECAF can output unstable levels of various disruption precursor, the data-driven property of our integrated framework allows it to be straightforward adapted for new devices or new

unstable events given sufficiently many new devices or new events data (i.e. retrain the model with new labeled data added to the original database) while the adaptation of physics-driven model to new operational region or unseen plasma instabilities requires physics understanding about the new physics and there does not exist a standard way to do this. Through extensive numerical experiments using data from the C-Mod, DIII-D, and EAST tokamaks, we demonstrate four major advantages of such an integrated framework:

- Any DL-based predictor can be adapted to an integrated model that combines event detector and disruption predictor using our framework with little extra computation.
- Numerical experiments show that the integrated model gives longer warning times for predicting disruptions when compared with a baseline disruption prediction model.
- The integrated model is able to identify the whole chain of events leading to the disruption instead of just predicting the final major disruption. The precursors' identification allows for implementation of appropriate actuators (a set of control knobs integrated in plasma control system, e.g. increase electron density or decrease plasma current.) than can be employed to actively avoid disruptions. Examples of control knobs that can be incorporated into the real-time plasma control system include increasing the electron density, or decreasing the plasma current.
- Finally, our cross-machine numerical experiments suggest that the combination of unstable events' identification with disruption prediction can strongly improve the cross-machine portability of the deep learning model.

5.2 Dataset description

Our disruption prediction and unstable event identification studies are conducted on disruption warning datasets coming from three experimental devices, i.e. Alcator C-Mod, DIII-D, and EAST [1]; additionally, a DIII-D dataset with labeled unstable events manually identified [19] is used. The three disruption warning datasets have been well described in our previous work [1, 10]. The dataset compositions and sampling rates of these three databases are shown in Table 5.1 [10]. We interpolate the signals from DIII-D and EAST onto uniform 10 ms and 25 ms time bases, respectively. This is necessary because the DIII-D and EAST disruption warning databases have nonuniform sampling for disruptive discharges [1, 20], while our DL-based model

Table 5.1: The dataset composition of the three disruption warning databases [10]

	No. shots (No. disruptive shots)	Sampling rate (ms)
C-Mod	4457 (932)	5
DIII-D	7105 (996)	10
EAST	11107 (3098)	25

requires uniformly sampled data. The manually labeled event identification dataset consists of 287 DIII-D disruptive shots (from the DIII-D 2015-2016 experimental campaigns), with manually labeled start times for different unstable events across the whole plasma current flattop of each shot [19]. We include 22 classes of different unstable events when we build this database, and all event names are consistent with events described in [1]. Given the limited size of the database and the frequency of different unstable events, we choose 10 classes of unstable events that occur during at least 10 different shots to include in our unstable event identification study Table 5.2. The *"event occurrence"* of a particular is the number of disruptive shots that have that event during the flattop, divided by the total number of disruptive discharges in the manually labeled DIII-D dataset (287). Since multiple unstable events can happen during the flattop of a single disruptive discharge, the sum of the "event occurrence" fractions can be larger than 1.

As for the selection of plasma signals included in our analysis, we first use all plasma signals considered in our previous disruption prediction study [10]. Furthermore, to better detect different unstable events, we add two more signals to the original list of plasma signals used by our model. The first additional plasma parameter is Te-width-norm, which is the half width of a parabola fitted to all measurement points from the core Thomson system, normalized by the minor radius. The core Thomson laser traverses the plasma vertically at fixed R, so our T_e fit is a function of vertical height, Z. The second additional plasma parameter is Prad-peaking-CVA, which is the radiation from the central plasma, divided by the total radiated power [21, 22]. The full list of input plasma signals is given in Table 5.3. The set of plasma signals included in this study is informed by three factors: 1. the suggestions from machine operators from C-Mod, DIII-D and EAST; 2. the analysis of the non-disruptive and close-to-disruption distributions of plasma signals included in our databases, as some signals have different distributions when disruption is imminent. For example, the normalized internal inductance, li, increases before the final current quench on C-Mod, DIII-D and EAST [1, 10, 20]; 3. We also take into account the need to characterize the plasma state and its evolution across the "events" or precursors considered for event identification. Plasma signals that are closely related to important disruption

Table 5.2: Event labels, descriptions, and occurrences for the manually labeled instabilities in the DIII-D dataset. Event labels follow [23]

Event label	Event description	Event occurrence
HL	H-to-L back transition	72%
ML	Mode locked	77%
RC	Radiative collapse	19%
MHD	$n = 1$ or 2 rotating MHD mode	61%
MAR	Multifaceted asymmetric radiation from the edge	7%
GWL	Greenwald density limit	5%
SAW	Large sawtooth crash	14%
IMP	Impurity influx	7%
IMC	Impurity control problem	5%
UFO	Unidentified impurity influx (flying macroscopic particles)	7%

precursors should be included in our analysis. For example, the n=1 locked mode is needed for detection of locked modes, which often precede disruptions.

The manually labeled database of DIII-D disruptive discharges is then randomly divided into a training set (160 shots) and a test set (127 shots). Our previous work [10] suggests that sequence-based models have a clear advantage over models based on individual time slice categorization. Therefore, we use plasma sequences of 10 consecutive time steps as input to our models. Since prior to each major disruption there is a sequence of unstable events that finally lead to the final loss of control, both the disruption prediction and the event detection problems are formalized as sequence-to-label supervised machine learning tasks. To this end, we need to assign two labels to each plasma sequence:

1. a disruption label, encoded as 1 if plasma sequences are close to disruption or 0 if the sequences are far from disruption;

2. a 10-dimensional event label vector, where each coordinate is independently linked to a score for one of the ten unstable precursors considered in Table 5.2. Each label vector element is encoded as 1 if the training plasma sequences are unstable with respect to the corresponding event, or 0 if the plasma sequences are stable.

Table 5.3: Plasma signals considered in the data-driven studies [10]

Signal description	Symbol
$\frac{\textbf{Plasma current}-\textbf{programmed plasma current}}{\textbf{Programmed plasma current}}$	`ip-error-fraction`
Perturbed field of nonrotating mode[a], $\frac{\mathbf{B^{n=1}}}{\mathbf{B_{tor}}}$	`locked-mode-proxy`
$\frac{\textbf{Electron density}}{\textbf{Greenwald density}}$	`Greenwald-fraction`
Distance between the plasma and the lower divertor	`lower-gap`
Current centroid vertical position error[b]	`z-error-proxy`
Plasma elongation	`kappa`
Poloidal beta	`betap`
$\frac{\textbf{Radiated power}}{\textbf{Input power}}$	`radiated-fraction`
Standard deviation of the magnetic field[c] **measured from an array of Mirnov coils, normalized by $\mathbf{B_{tor}}$**	`rotating-mode-proxy`
Loop Voltage $\mathbf{V_{loop}}$	`v-loop`
Normalized internal inductance	`li`
Safety factor at 95% flux surface	`q95`
Fitted half width of the $\mathbf{T_e}$ profile from Thomson scattering normalized by minor radius	T_e`-width-norm`
Radiation from central plasma divided by the overall plasma radiation	P_{rad}`-peaking-CVA` [21, 22]

[a]For the C-Mod database, the `locked-mode-proxy` signal is obtained from a Mirnov coil array instead of the saddle coil.
[b]For the DIII-D database, we use current centroid vertical position instead of position error for the `z-error-proxy signal`.
[c]For the DIII-D database, we use n=1 component of magnetic field measured from a Mirnov coil array normalized by B_{tor} for the `rotating-mode-proxy` signal.

For disruption label assignment, we use the same procedure as our previous study [10]. However, for event label assignment, the procedure is not straightforward: (1) We only record the start time of each unstable event in our manually labeled dataset, but the end time of the event is missing. (2) All manually labeled shots are disruptive shots, but we need both disruptive and non-disruptive training shots for the development of our integrated model. In the following we present the solutions to both these problems.

After testing different label assignment schemes, we find that the best approach is to label all plasma sequences that encompass the start time (onset point) of the unstable events as belonging to the unstable category of the corresponding event. All other plasma sequences that are either before or after the onset time belong to the stable category of the corresponding event. Under this labeling scheme, our target is to identify the onset of unstable events instead of the unstable events themselves. The predicted onset time is the point at which (i) the predicted event's level exceeds the threshold corresponding to the unstable event and (ii) this event's level is larger than the level of the event on the previous time step.

Finally, in order to complement our disruptive dataset of labeled unstable events, we randomly select 900 non-disruptive shots from the 2015-2018 DIII-D experimental campaigns and assign unstable event labels to these 900 non-disruptive shots using a trained event predictor through an iterative labeling process that will be discussed in detail in Section 5.3.

By solving these two problems, we construct a database with events and disruption labels for both disruptive (manually labeled, 160) and non-disruptive (automatically labeled, 900) shots that represent the training set for the development of the integrated DL model. Notice that 160/900 is close to the ratio of disruptive and non-disruptive discharges in our DIII-D disruption warning database [1, 10]. In addition, 700 non-disruptive shots randomly selected from the DIII-D disruption warning database are combined with 127 manually labeled disruptive shots to form a test set. Both disruptive and non-disruptive test data are used to evaluate disruption prediction performance of the model while only disruptive test data are used for testing of disruption precursor detection performance of the model. Finally, 127 disruptive shots and 700 non-disruptive shots are randomly selected from the DIII-D disruption warning database as the validation set. For the disruption prediction problem, the time threshold that determines the unstable phase of each disruptive training sequence (described in [10]) is uniquely chosen as the time at which the first unstable precursor event appears.

The training samples are $(x, (y_{dis}, y_{event}))$ pairs where x is a 10-step consecutive temporal sequence of the 14 plasma signals in Table 5.3, and y_{dis}, y_{event} are the disruption label and event label, respectively. The training samples are extracted

from each training set via a scheme equivalent to that used in reference [10]. For each disruptive training shot, 20 disruptive samples are randomly selected from those sequences that intersect the unstable phase of the shot. For each non-disruptive training shot, 20 non-disruptive samples are randomly selected from all sequences during the flattop phase of the plasma current.

5.3 Labeling non-disruptive shots through an iterative labeling process

As mentioned in Section 5.2, assigning event labels to non-disruptive data in the training set is necessary for the development of the integrated model. A previous study [19] gives examples of using data-driven methods to generate event labels for unseen shots, given a very limited number of manually labeled shots. Given a manually labeled dataset with 300 shots, we want to develop an event identifier to automatically assign event labels to non-disruptive training shots. To this end, we designed an event identifier and used a trained event identifier to generate event labels for all 900 non-disruptive discharges in the training set via an iterative labelling process. We note that the size of the manually labeled dataset is relatively small, and it only includes disruptive shots from DIII-D 2015-2016 campaigns. Given this limitation, the distribution of different events in this dataset might be incomplete or biased and it might miss some event chains that can lead to disruption. Generating event labels for non-disruptive discharges, using a model trained on this manually labeled dataset, can result in biased event labels because the trained model will be affected by the event occurrence in the training set, and it can only recognize those patterns of unstable events that appear in the training set. Therefore, the event identification performance of the final trained model on manually labeled dataset can be exaggerated. Nevertheless, adding automatically labeled data to the training set should still improve the event detection performance of the trained model as long as the generated labels are accurate enough. In addition, a larger training set provides higher statistical significance. Furthermore, since the disruption label for each training shot is already known, and disruption labels are independent from event labels, the biased event labels should only have small effects on the disruption prediction results. This is because disruption prediction does not require us to detect **all** events, but rather only those **typical/frequent** events in the event chains that lead to disruptions. As long as our biased event dataset covers the most frequent unstable events that lead to disruptions (e.g. HL, ML, MHD on DIII-D), the trained event identifier should be able to detect these frequent events and give us extra warning time. The biased event labels, and hence event identifiers, might miss some infrequent events, but these

missed events should have a negligible effect on disruption prediction. In this section, the event identifier and iterative labeling process are discussed in detail.

5.3.1 The Hybrid Deep Learning (HDL) event identifier

Modified with respect to our previous work [10, 24], a hybrid deep learning (HDL) model is developed for unstable event identification. The HDL event identifier (HDL-EI) consists of six multi-scale temporal convolution (MSTConv) layers and two dense-bn layers, plus input and classification layers (with sigmoid activation [25] for each coordinate). A dense-bn layer (Figure 5-1(c)) contains a fully connected layer followed by a batch normalization layer [26] and a rectified linear unit (ReLU) activation [27]. The MSTConv layer, described in [10], is a novel neural network layer designed for time-series processing. It contains six 1D temporal convolution layers, as well as batch normalization and ReLU activation. The architecture of the HDL-EI is shown in Figure 5-1(a) and the structure of the MSTConv layer is detailed in Figure 5-1(b). Empirically, the deep neural network is designed to have wider layers in the middle of the model, which allows the network to identify more complex patterns within the input data. Accordingly, the third, fourth and fifth MSTConv layers, in the middle of the neural network, have 15 convolutional filters in each of their 1D temporal convolution layers, while the first, second and sixth MSTConv layers have 10 convolutional filters in each of their 1D temporal convolution layers. The wider MSTConv layers in the middle of the neural network allow for more complex intermediate representation. This architecture gives better performance than the model that uses 10 filters for each MSTConv layer.

The HDL-EI transforms an input 10-step consecutive temporal sequence of 14 plasma signals to an output 10D event level vector at the last time step of the sequence. Each coordinate of the event level vector provides unstable levels for one event in Table 5.2, with ranges between 0 and 1, where 1 is the unstable class and 0 is the stable class; the training loss of the HDL-EI comes from an average mean square error (MSE) of each individual unstable event. To label non-disruptive data, each shot was divided into batches of sequences, with each neighboring sequence having 9 steps of overlap. Therefore, given a non-disruptive shot, with N flattop time steps from t_1 to t_N, the HDL-EI will generate N-9 event level vectors, corresponding to the time steps between t_{10} and t_N. If one coordinate (e.g. the third coordinate) of the output event vector exceeds the preset threshold corresponding to that event (e.g. HL, 0.5) at a flattop time step, while it is less than the event threshold at previous time step, the time of this step is identified as the predicted onset time $t^i_{onset,p)}$ of the corresponding event (e.g. HL) and the $t^i_{onset,t)}$ means the manually labeled onset time of the corresponding event.. A simple illustration of this process is shown in

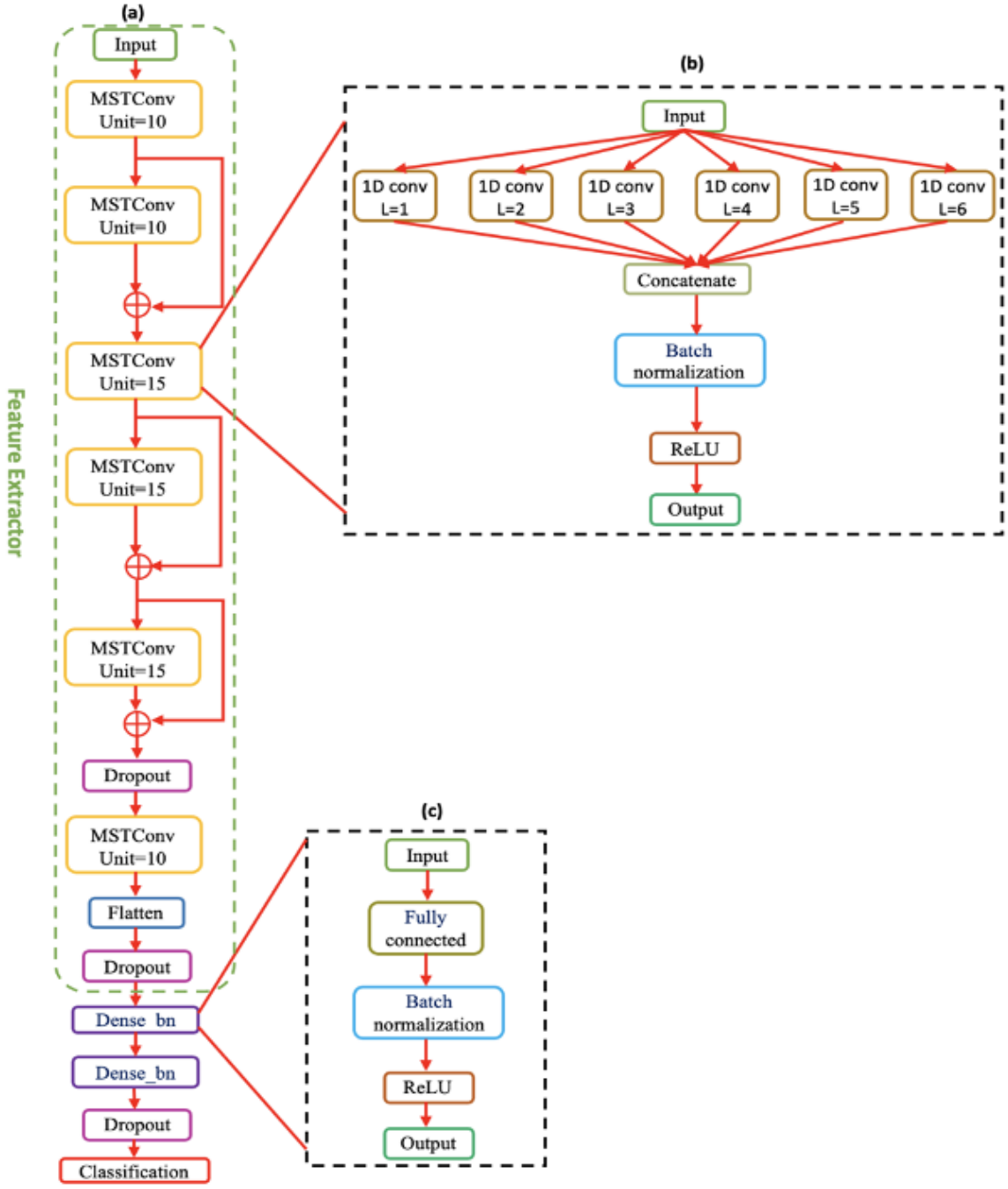

Figure 5-1: (a) The HDL-EI; (b) the detailed structures of the dense-bn layer; and (c) the MSTConv layer. The feature extractor of the HDL-EI is marked by a green dashed box. Note that the six 1D temporal convolution layers contained in the MSTConv layer have window lengths L from one to six, to extract local temporal information at the different levels (see [10] for a detailed explanation).

Figure 5-2. To evaluate the shot-by-shot performance of the HDL event identifier, we focus on the first onset of each unstable event during the test shot. If the predicted first onset time is close (within uncertainty) to the manually labeled first onset time: $|t^i_{onset,t} - t^i_{onset,p}| < 0.03s$, then it is considered a true positive. Different thresholds were considered and 30 ms represents the best trade-off, allowing us to achieve good average accuracy (above 80%) for the five most frequent events (HL, ML, RC, MHD, SAW). Furthermore, we find the class membership probabilities for each particular event (aka instability levels from HDL-EI) corresponding to these five events usually ramp up within 30 ms of unstable event onset .These observations suggest that 30 ms is a good choice for the definition of the TP criterion for these five most frequent events, and a 30 ms time interval is a good match to the time scale of these five events on DIII-D. If the output event level corresponding to an event does not exceed the threshold for the whole flattop of a shot, and this event does not happen during the flattop of this shot, this is regarded as a true negative. HDL-EI is optimized to achieve the highest TPR at a fixed FPR (typically FPR=0.1). From Table 5.2, it is clear that, from among the ten selected events, HL, ML, RC, MHD, and SAW have the highest frequencies. To maximize the overall accuracy of the model, a good model should give higher weight to the TPR in the ML detection (since the occurrence of ML is 77%) to avoid missed alarms, and while giving more weight to the FPR for IMC detection (since the IMC probability is low) to avoid false alarms. Due to these considerations, when we define the performance metric for each event, we choose different target FPRs for frequent and infrequent events, allowing us to rebalance the class frequencies for different events.

5.3.2 Iterative labeling process

The iterative labeling process evolves in two stages. During the first stage, the initial training set, X_1 of the HDL-EI, is constructed by sampling 20 sequences (each sequence is a 10x14 matrix) from the unstable phase of each manually labeled disruptive training shot. The HDL-EI is then trained using this initial training set. The trained model is subsequently applied to manually labeled disruptive test shots, and the optimal threshold corresponding to each event is obtained by optimizing the performance of the HDL-EI on the manually labeled test set. The performance of HDL-EI model in the first stage is given in Table 5.4. After this, the predicted event labels of all 900 non-disruptive shots are generated using the trained model and optimized event thresholds. During the second stage, the training set is obtained by sampling 20 sequences from the unstable phase of each manually labeled disruptive training shot, plus randomly sampling 20 sequences from the flattop of each non-disruptive training shot (with generated event labels). The HDL-EI is then trained

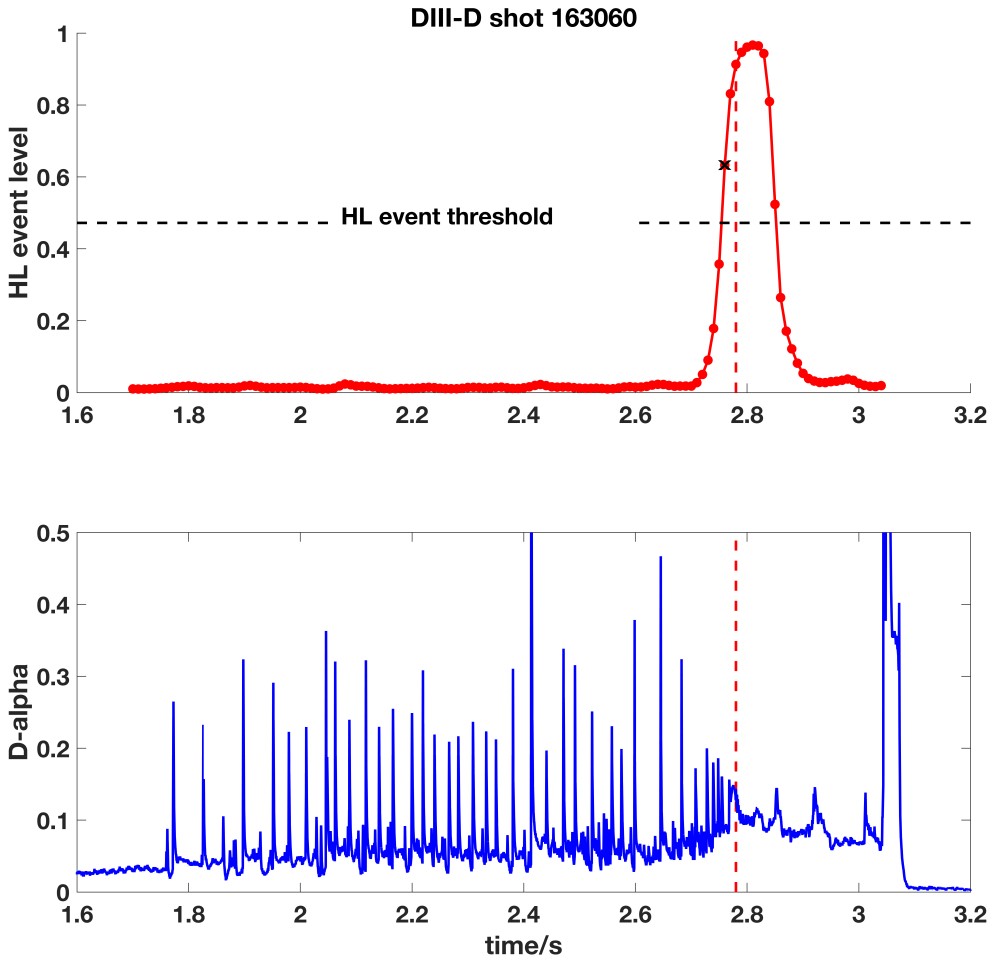

Figure 5-2: The upper panel shows the output HL back transition level from the trained HDL-EI. The manually labeled HL onset time is marked by the vertical dashed line, and the preset HL back transition event threshold is marked by the horizontal dashed line. The predicted onset time ($t^1_{onset,p} = 2.76\,\mathrm{s}$) of HL is marked by a black X on the output HL level. At this time step, the output HL level (0.633) is greater than the threshold, while the output HL level at the previous time step (0.357) is below the threshold. The time trace of a D_α signalis shown in the lower panel. The large spikes correspond to type I ELMs; the last ELM occurs just prior to the HL back transition.

Table 5.4: HDL-EI performance on test set in stage 1 of the iterative labelling process

Frequent events (FPR= 0.15)	TPR	Infrequent events (FPR= 0.10)	TPR
HL	0.80	MAR	0.75
ML	0.85	GWL	1.00
RC	0.72	IMP	0.33
MHD	0.67	IMC	0.57
SAW	0.59	UFO	0.40

using this combined training set. Given the trained model, the optimized threshold corresponding to each event, and the predicted label of each non-disruptive training shot, are obtained using the same method as in stage 1. The second stage of the labeling process is run iteratively until the obtained thresholds and the performance on the manually labeled test set converge. The ensemble method is well known in the machine learning community, and has been shown to significantly increase the performance and reduce the uncertainty of the model [10, 28–30]. In our previous work [10], we have shown that using the ensemble method can significantly improve the performance of a data-driven disruption predictor. Therefore, we independently trained 10 different HDL-EIs with the same dataset; each HDL-EI has different initial parameters (i.e. different initialization) and different training random seeds. Then, we combine these 10 independently trained HDL-EIs into an ensemble. The final output of our model is the average output from ensemble of 10 HDL-EIs. The final event thresholds and generated event label for non-disruptive training shots are obtained from the average output of an ensemble of ten independently trained HDL-EIs. The diagram of this iterative labeling process is shown in Figure 5-3, and the final optimized event thresholds are summarized in Table 5.5. Notice that infrequent events tend to have lower thresholds. This phenomenon comes from the fact that the HDL-EI always sees negative samples during training and it learns to always output a low event level to achieve high accuracy. The low event level leads to a low event level threshold. An example of automatically labeled non-disruptive training shots is shown in Figure 5-4.

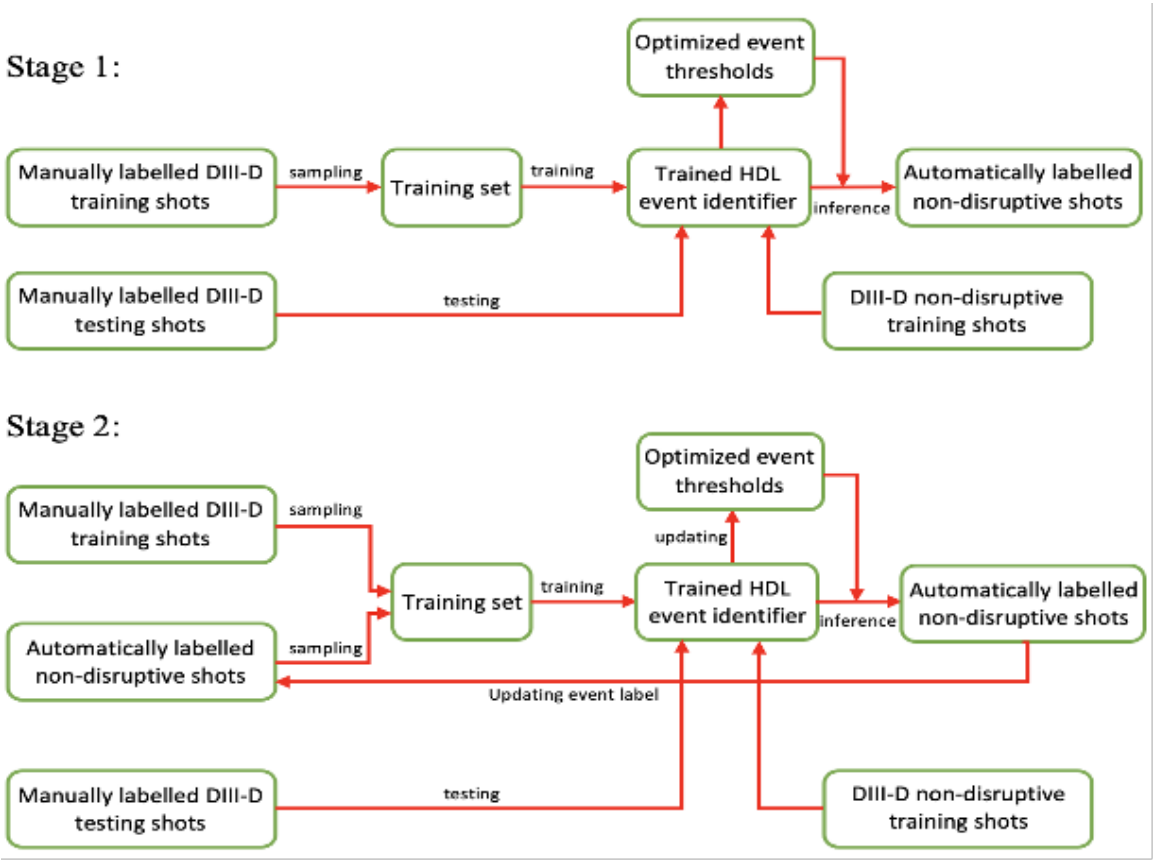

Figure 5-3: The diagram of iterative labeling process.

Table 5.5: The optimized event thresholds from iterative labeling process (See Table 5.2 for event descriptions.)

Event	Threshold	Event	Threshold
HL	0.472	GWL	0.091
ML	0.655	SAW	0.151
RC	0.262	IMP	0.057
MHD	0.570	IMC	0.034
MAR	0.023	UFO	0.053

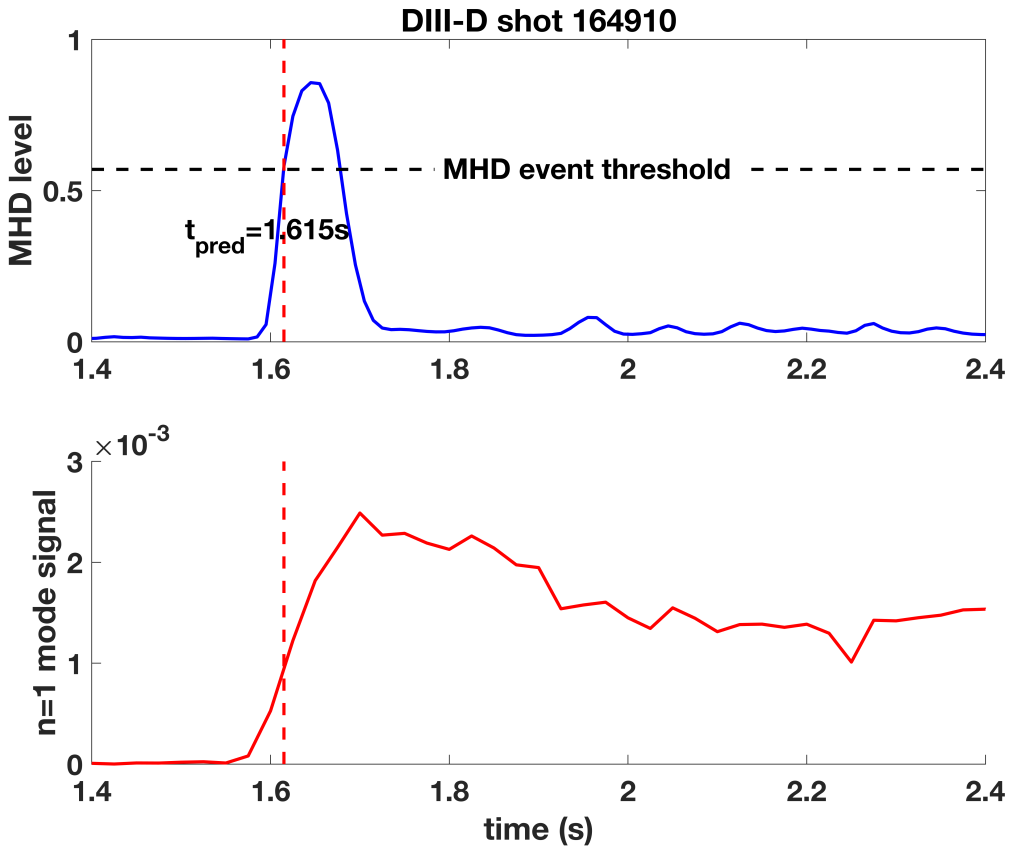

Figure 5-4: An example of an automatically labeled non-disruptive shot from DIII-D. For this shot, the large 2/1 tearing mode happens at 1.60 s, shortly after the plasma enters H-mode. The $n = 1$ tearing mode onset time is marked as a vertical line in the plot, and the predicted MHD level exceeds the threshold 15 ms after the onset. Notice that only the predicted label of the MHD event is shown in the plot, as all other event levels are close to 0 and do not exceed the corresponding event thresholds.

5.4 The integrated deep learning framework for disruption prediction and unstable event identification

The integrated DL framework combining the predictive ability of disruptions, as well as several precursors, is developed using the training set that includes manually labeled shots and automatically labeled ones. This integrated framework is designed to map an input plasma sequence to two connected outputs: a scalar indicating the disruption risk, and a 10-D event level vector that corresponds to the level of all 10 classes of unstable events. The model's loss function includes two terms that need to be minimized at the same time. Figure 5-5 shows the architectural details of this deep learning framework. Since the disruption level is closely related to the unstable levels of each disruption precursor, we want the intermediate representation of the input signals to contain information about both the precursors and the major disruption itself. The integrated model is built on the HDL-EI described in Section 5.3 by adding a separate disruption prediction branch after the intermediate layer of the original HDL-EI. This allows the model to output both the disruption level, i.e. the "disruptivity", and the predicted event level vector based on the intermediate representation of the input plasma signals. The integrated model adopts a composite loss function (a function that measures the difference between predicted label and ground truth) that includes the contributions from both the unstable event identification and the disruption prediction branch. This loss function can be represented as:

$$loss_{integrated\ model} = loss_{dis} + \lambda * loss_{event} \tag{5.1}$$

where $loss_{event}$ is the average mean squared error (MSE) loss of the unstable event task, while $loss_{dis}$ is the average negative log-likelihood (NLL) loss of the predicted disruptivity risk. λ is the framework's hyperparameter balancing these two terms, and we have chosen $\lambda = 1$ for these studies. By removing the event branch of the integrated model, and setting $\lambda = 0$ in Equation (5.1), we can convert an integrated DL model into a baseline HDL disruption predictor.

The shot-by-shot testing scheme of the integrated framework follows the two-staged approach of the HDL-EI iterative labeling. If the level of any unstable event (e.g. ML) or of the disruptivity exceeds the corresponding preset threshold at any flattop time step, the whole shot will be classified as unstable (with respect to that event, e.g. ML) or as disruptive shot. A successfully predicted DIII-D disruptive shot from the test set is shown in Figure 5-6, and the event identification performance of the model is given in Table 5.6. The average TPR of the four most frequent unstable

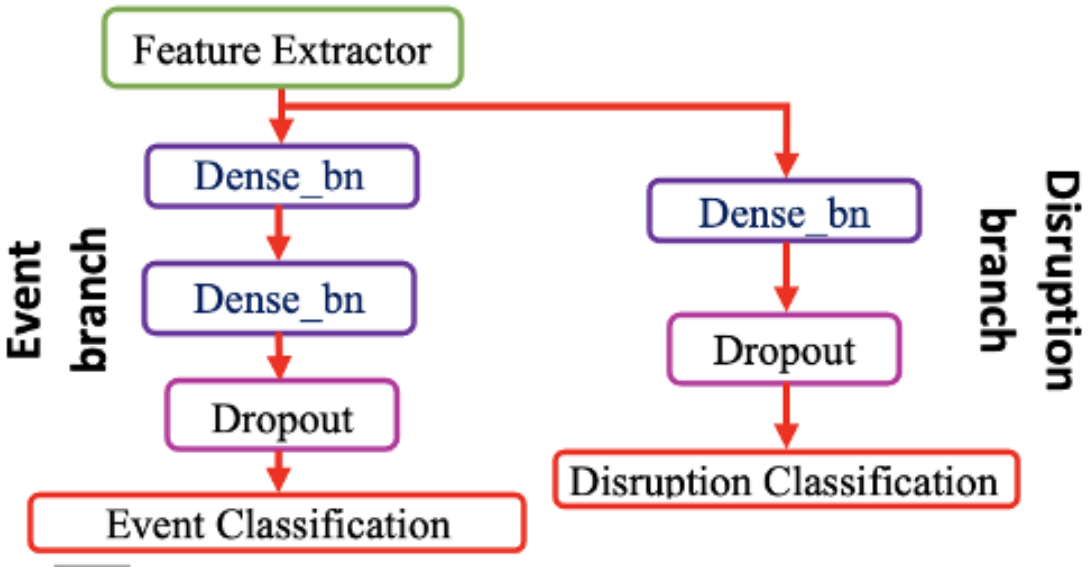

Figure 5-5: The architecture of the integrated deep learning framework. The detailed structure of the feature extractor is given in Figure 5-1.

events (HL, ML, RC, MHD) achieves 84%, which is significantly better than the performance of HDL-EI when only trained with manually labeled data (see Table 5.4); this confirms the effectiveness of using automatically generated event labels.

5.4.1 Comparing the disruption prediction performance between the integrated model and the baseline predictor

To investigate the advantage of an integrated DL model, we compare the performance of the integrated model with that of the baseline HDL disruption predictor, using the same test set for each approach. Both the integrated model and the baseline model

Table 5.6: Event identification performance of the integrated DL model on manually labeled DIII-D test shots

Frequent events (FPR= 0.15)	TPR	Infrequent events (FPR= 0.10)	TPR
HL	0.89	MAR	1.00
ML	0.92	GWL	1.00
RC	0.81	IMP	0.33
MHD	0.73	IMC	0.71
SAW	0.59	UFO	0.60

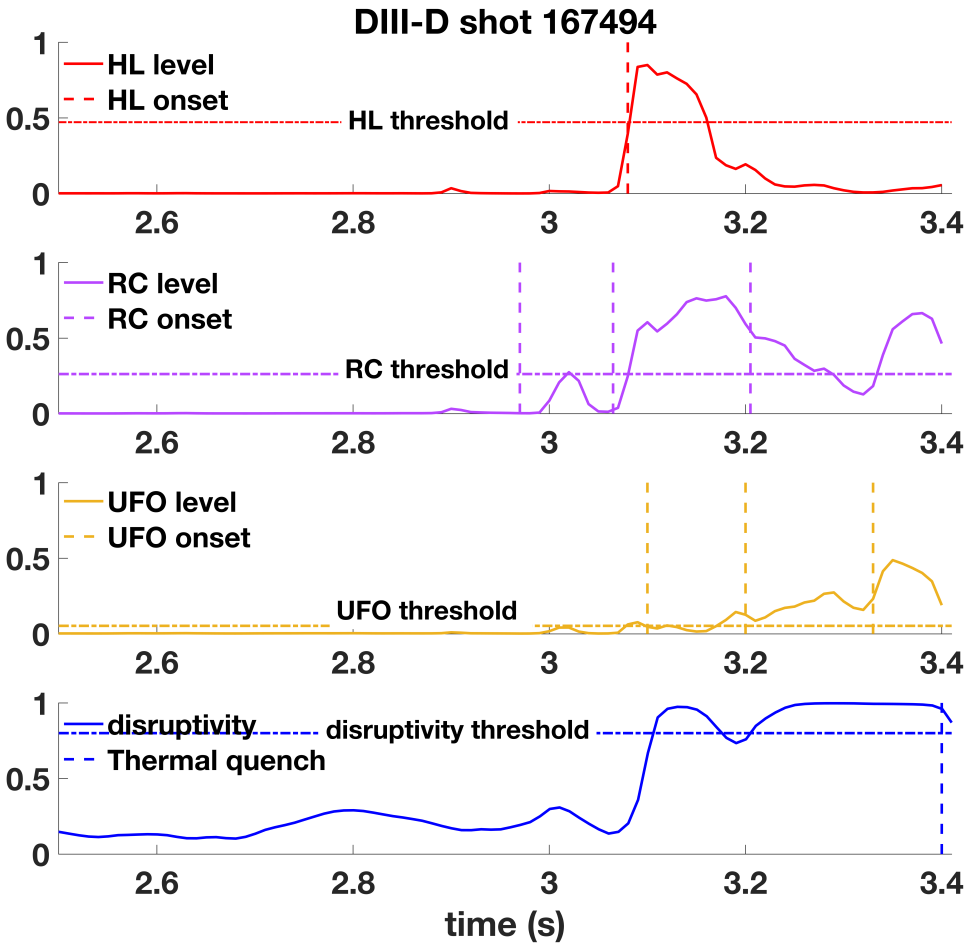

Figure 5-6: A successfully predicted DIII-D disruptive shot from the test set. All the time traces corresponding to the events that pass the event threshold are shown in the plots with solid lines, and the manually labeled onset time of each unstable event is also given in the plot as a vertical line with the same color as the corresponding event level line. The event thresholds are marked as horizontal lines.

are trained using the same DIII-D training set. The baseline model doesn't need event labels from the training shots; it only uses the threshold for major disruption (disruptivity threshold). The event thresholds are obtained via an iterative labeling process, and they are fixed during this experiment. The performance metric chosen for these numerical experiments is the area under the receiver-operator characteristic (ROC) curve (AUC), which is the curve of true positive rate (TPR, the ratio of correctly predicted disruptive shots to all disruptive shots) and false positive rate (FPR, the false alarm rate) [31].The disruption prediction performances of all numerical experiments reported here are evaluated at 50 ms before the current quench, as this is the requisite warning time needed to successfully trigger the mitigation system on ITER [32]. The comparison results are shown in Figure 5-7. To ensure a fair comparison, the hyperparameters of each disruption predictor are optimized independently using a separate validation set. To do this, we independently tune the hyperparameters of the baseline HDL model and the integrated DL model, to maximize their disruption prediction performance on this validation set. In addition, for true positive shots, the cumulative distribution of warning times, (i.e. the difference between the triggered alarm time t_{alarm} and the disruption time t_{dis}), returned by the two models, are reported in Figure 5-8. Through the comparison, the integrated DL model gives AUC=0.940 (TPR=0.88 at FPR=0.1) while the baseline HDL model gives AUC=0.920 (TPR=0.85 with FPR=0.1). Note that the 0.940 AUC achieved by integrated model is close to the performance of the original HDL model, trained on a much larger dataset (reported in [10]). There are three major factors that contribute to this: 1. Adding event information; 2. Improved network design, substituting a GRU layer with an MSTConv layer and adding short-cut connection; 3. Adding two useful 1D features (Te-width-norm and Prad-peaking-CVA). If we consider the fact that the baseline predictor has already achieved high accuracy on DIII-D, the 3% TPR improvement is significant. By using the integrated DL model, we reduce the number of the missed alarms by 20% (15 missed alarms to 12 missed alarms every 100 disruptions). The integrated DL model also gives longer median warning times compared with the baseline disruption predictor. The median warning time increases by roughly 200 ms when we use the integrated DL model, and the longer warning time could allow the plasma control system to take actions to avoid disruptions, instead of simply mitigating them. Furthermore, detecting unstable events together with disruption in plasma experiment operation enables disruption avoidance and analysis of plasma physics. All these considerations contribute to clarifying the advantage of the integrated HDL vs the baseline version.

The conclusions above suggest the advantage of providing unstable event information to such DL frameworks and verify the close correlation between unstable event identification and disruption prediction tasks. In Figure 5-8, most of DIII-D shots

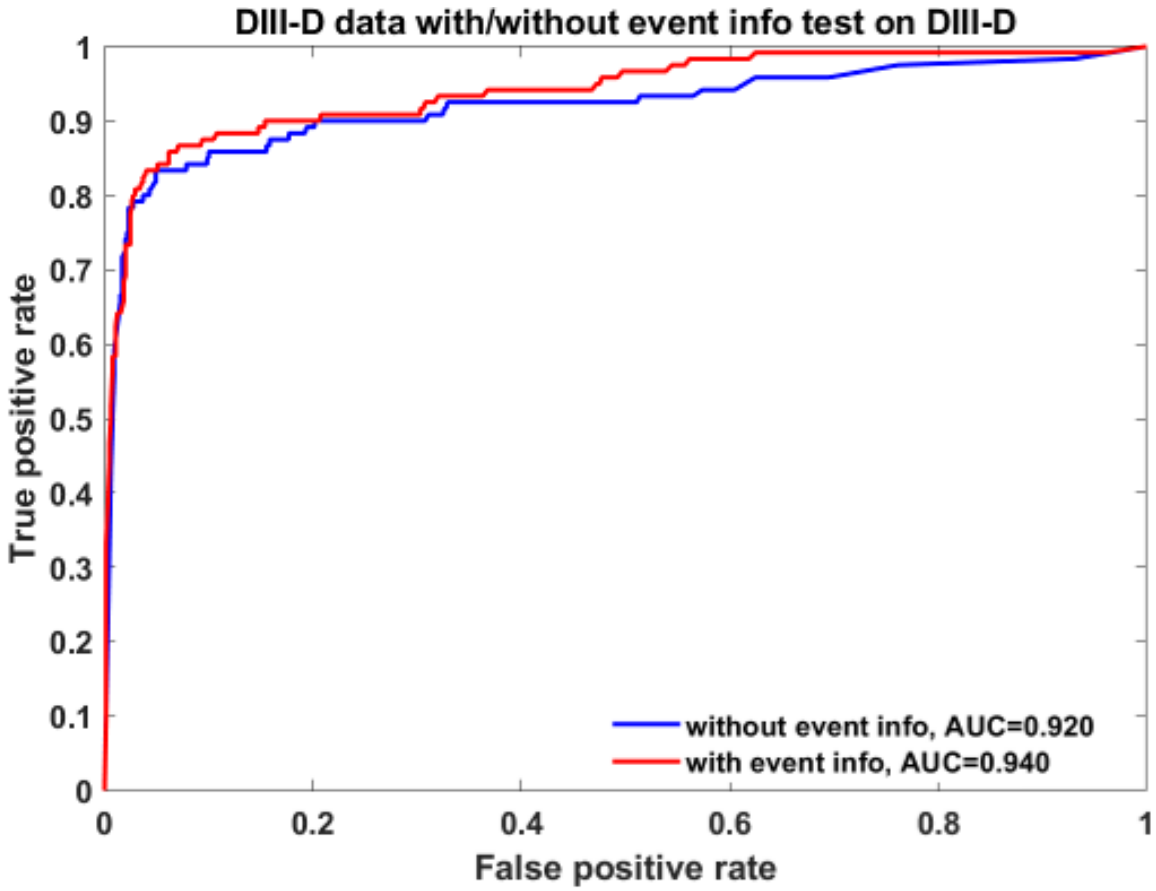

Figure 5-7: The ROC curves from DIII-D test sets for the integrated deep learning model (using event information, red) and for the baseline disruption predictor (without event information, blue).

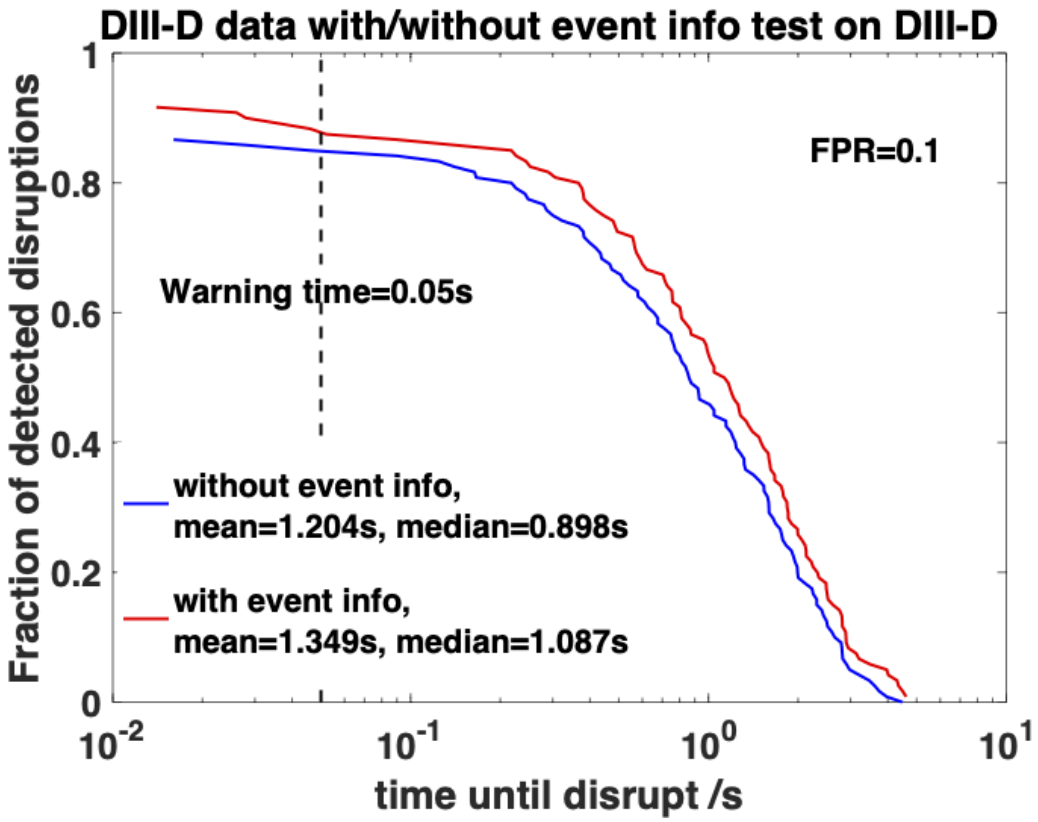

Figure 5-8: The cumulative distributions of warning time from DIII-D test sets returned by integrated deep learning model (using event information, red) and baseline disruption predictor (without event information, blue). The vertical dashed line shows the 50 ms warning time threshold.

that have very long warning time greater than one second usually have locked mode onset at the early/middle stage of the flattop but the initial locked mode onset does not result in the large thermal quench of the plasma in next few hundred milliseconds (either survive with locked modes or locked modes disappear after onset).

Having demonstrated that adding event information helps the integrated model achieve higher accuracy on disruption prediction, we ask a further question: to improve the disruption prediction, how accurate does the unstable event identifier need to be? To try to answer this question, we reduced the size of manually labeled training set (from 160 shots to 110 shots) and used an iterative labeling process on this reduced training set to label 50 remaining disruptive shots and 900 non-disruptive shots. Then, we combine these 50 disruptive shots, plus 900 non-disruptive shots, with generated labels and 110 manually labeled shots to the new "degraded" training set. Finally, we train a "degraded" integrated DL mode using this new combined dataset. The event identification performance of the "degraded" integrated DL model is shown in Table 5.7; the average TPR for the four most frequent unstable events (HL, ML, RC, MHD) is 0.74. The disruption prediction performance of the degraded integrated DL model, and the comparison with both the complete integrated DL model (trained with all event information) and with the baseline HDL model, are given in the Table 5.8. From the comparison, the "degraded" integrated DL model gives similar disruption prediction performance compared with baseline HDL model, which suggests that a bad event identifier might not be able to provide extra information for disruption prediction. Results from Table 5.8 show that the event identifier needs to achieve higher than 75-80% accuracy for the most frequent unstable events in order to improve the disruption prediction. We need to mention that the 75-80% accuracy estimation is not directly applicable to other devices, because different devices have different frequent events and different event occurrence. This empirical accuracy should also depend on the signals considered by model and the accuracy of the baseline model. The required accuracy will decrease if the baseline model performance is lower. Knowing the statistics of the root cause of disruptions on the tokamak [23, 33] might help us obtain an upper bound of this required accuracy. However, since this value depend on lots of factors, more accurate estimation of the required accuracy needs to be obtained via numerical experiments.

5.5 Cross-machine performance of the integrated model

Given the fact that a few, or even one unmitigated full current, high stored energy disruption can significantly damage future tokamaks, including ITER, it is strongly desirable to develop a disruption predictor that can reliably and accurately operate before the first high performance operation of the device [34]. Therefore, the disrup-

Table 5.7: Event identification performance of the "degraded" integrated DL model

Frequent events (FPR= 0.15)	TPR	Infrequent events (FPR= 0.10)	TPR
HL	0.78	MAR	0.50
ML	0.82	GWL	1.00
RC	0.69	IMP	0.33
MHD	0.67	IMC	0.43
SAW	0.47	UFO	0.60

Table 5.8: The performance of the integrated DL model trained with "degraded" event information

	AUC	TPR at FPR=0.10	Median warning time (s)
Baseline HDL	0.920	0.852	0.898
Degraded integrated DL model	0.916	0.849	0.906
Complete integrated DL model	0.940	0.883	1.087

tion prediction model with better cross-machine transferability represents a suitable candidate for the DMS trigger algorithm for future devices like ITER, assuming that enough knowledge from other tokamaks' data is extracted, and that only a minumal amount of data from the new machine itself is required. From Section 5.4, we find unstable event information can provide extra information and improve the disruption prediction performance of the data-driven model. Next, we would like to investigate the cross-machine transferability of the integrated framework by setting up extensive numerical experiments, and comparing disruption prediction performances against the baseline DL model.

In this section, we consider DIII-D as the "*existing*" device with C-Mod or EAST chosen as the "*new*" device and investigate how the integrated model (with event information) and baseline disruption predictor (without event information) trained on DIII-D data perform on either C-Mod or EAST. The description of the C-Mod and EAST disruption warning databases can be found in [1, 10].

5.5.1 Cross-machine prediction performance of the integrated model and baseline disruption predictor

The cross-machine transferability of the following models is considered in the comparison experiments:

- The integrated model trained on the combined DIII-D dataset, including the event information detailed in Section 5.3 – i.e., all disruptive training shots have manual event labels and all non-disruptive training shots have generated event labels, and
- The baseline disruption predictor trained on the exact same training data, but without event information – i.e., removing all event labels.

The trained integrated model and the baseline model are tested on the EAST and C-Mod datasets, respectively. Besides the prediction accuracy, we are also interested in investigating whether providing unstable event information from the "*existing*" device allows the trained model to find early precursors of the disruption on a different, "*new*" machine, and hence give longer warning times. To this end, the distributions of warning times returned by the integrated model and the baseline disruption predictor are also analyzed. Results on the test set are shown in Figure 5-9 and Figure 5-10. From these test results, we draw the following conclusions:

- Comparing the ROCs of the integrated DL model (red) and the baseline HDL model, the integrated DL gives TPR= 0.61 at FPR= 0.20 for EAST, and baseline HDL gives TPR= 0.50 at FPR= 0.20 for EAST. The integrated model gives better cross-machine accuracy compared to the baseline disruption predictor (Figure 5-9(a), Figure 5-10(a)).
- The integrated model provides longer warning times compared to the baseline disruption predictor, even under the cross-machine scheme. When the FPR is set to 0.2, the integrated DL gives 120 ms median warning time on EAST, while the baseline HDL gives 50 ms median warning time on EAST (Figure 5-9(b), Figure 5-10(b)).

These two conclusions imply that information about unstable events, i.e. disruption precursors, contains general, machine-independent knowledge about disruptions. The common physics contained in the event information can be learned by an integrated model, and gives better transferability across devices when predicting disruptions.

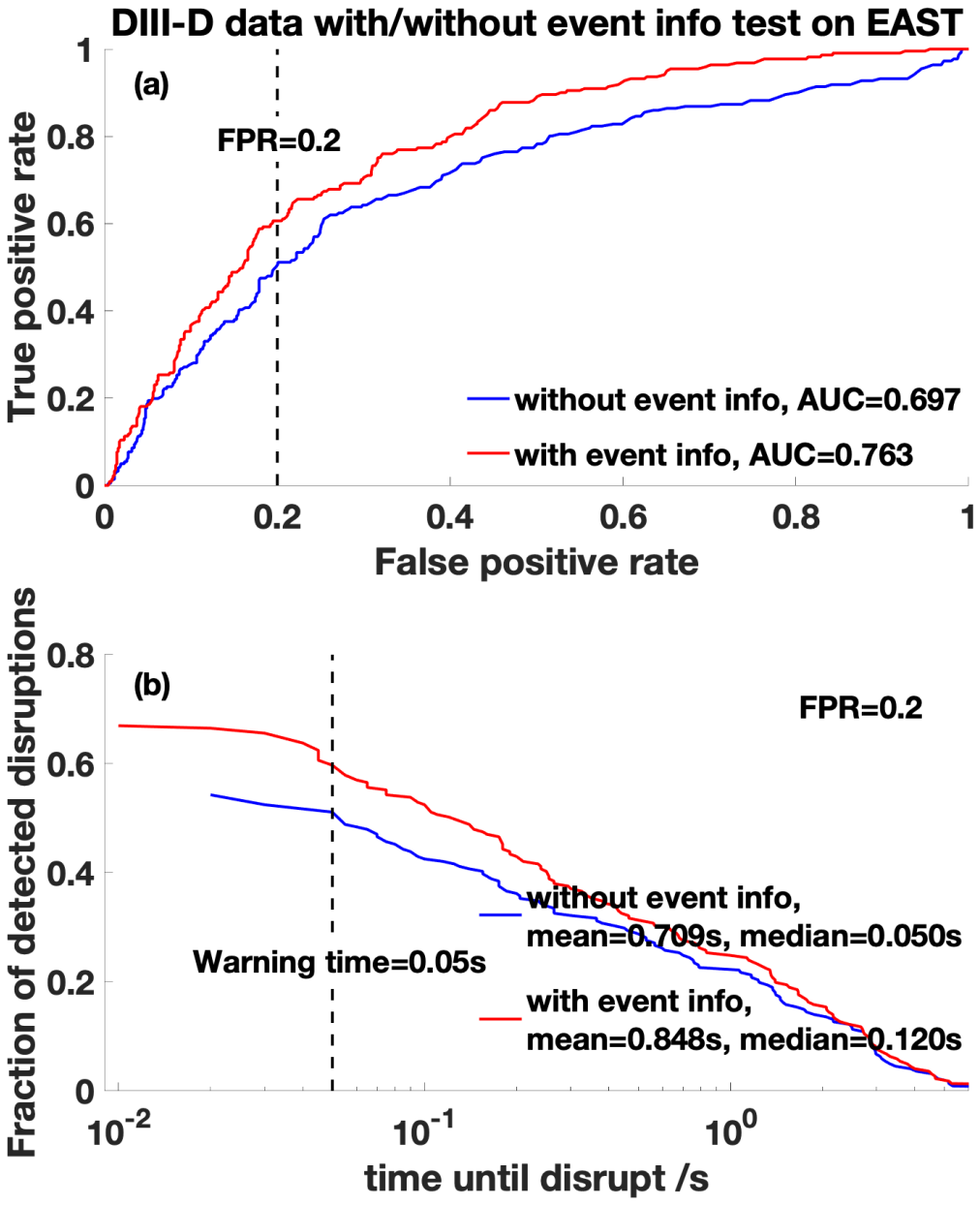

Figure 5-9: The ROC curves from the EAST test set for the integrated deep learning model (using event information, red) and for the baseline disruption predictor (without event information, blue) trained on DIII-D training set. The vertical dashed line in the upper panel corresponds to FPR=0.2 and the vertical dashed line in the lower panel shows 50ms warning time threshold.

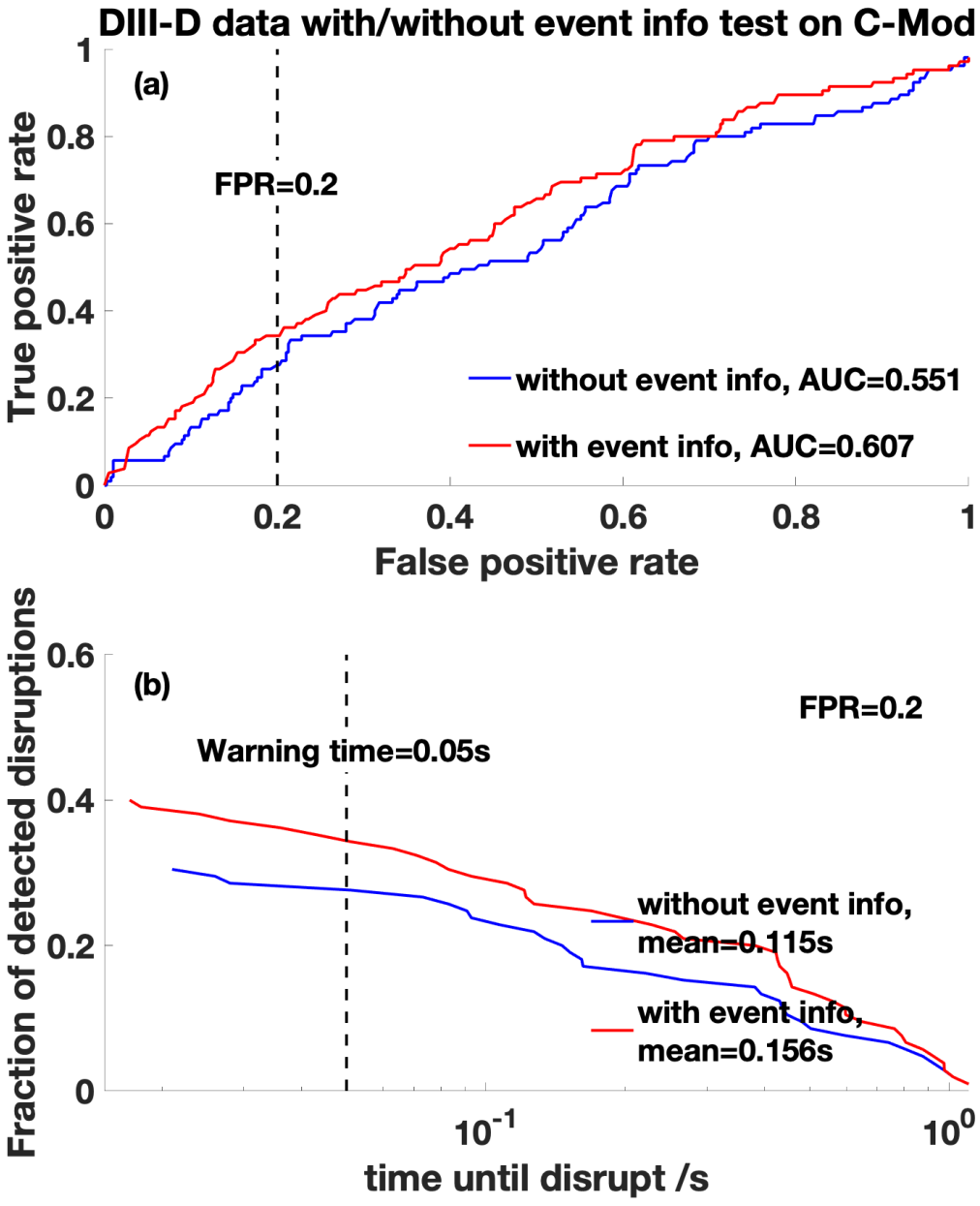

Figure 5-10: The ROC curves from the C-Mod test set for the integrated deep learning model (using event information, red) and for the baseline disruption predictor (without event information, blue) trained on the DIII-D training set. The vertical dashed line in the upper panel corresponds to FPR=0.2 and the vertical dashed line in the lower panel shows 50ms warning time threshold.

5.5.2 Cross-machine unstable event identification

Beyond the cross-machine disruption prediction transferability, we also want to investigate whether the integrated model, trained with data from one machine, can identify disruption precursors on different devices. To this end, we manually labeled the unstable events of a few shots from EAST and C-Mod, and applied our integrated model trained on DIII-D labeled shots. While the number of manually labelled EAST and C-Mod test shots is small, we can still get preliminary cross-machine conclusions from the test. The test results are shown in Figure 5-11 and Figure 5-12. The results from these experiments point to the following qualitative conclusions:

- The integrated model trained with data from one device, can qualitatively identify disruption precursors on different tokamaks. This observation again suggests that the underlying physics driving the unstable events is similar across tokamaks (Figure 5-12 and Figure 5-11).

- The integrated model trained on data only from one device seems to have large numerical bias when it is directly applied to another device. This bias can make the absolute values of disruptivity and unstable event level returned by model meaningless. Nevertheless, the increases of the output instability levels from the model still indicate the increasing risks of correspondent plasma instabilities (Figure 5-11). The underlying reason for this cross-machine bias is that different machines have very different operational regimes in the parameter space, even after signal normalization [10]. Therefore, the data-driven model trained on data from device A can be unconstrained on a disjoint new operational regime (e.g. the operational regime of a different tokamak, B) because the model has never seen a sample related to this new regime. The extrapolation of the trained model to this new regime can result in large numerical bias. This numerical bias can be greatly reduced by adding a few shots from the target domain/device [9, 10, 24] and/or by adding some simulation data to the training set. Previous studies [1, 10, 20, 24] have shown that different devices usually have similar behaviors when a disruption/unstable event is imminent. These similar dynamics among devices can be captured by data-driven models (especially sequence-based models) and hence the changes of the output instability levels from the data-driven model trained on other devices can still reflect the risks of corresponding unstable events [10].

5.5.3 Summary of Cross-machine numerical experiments

Given all the conclusions in Section 5.5.1 and Section 5.5.2, it is possible to state that disruption prediction and unstable event identification are two closely related tasks

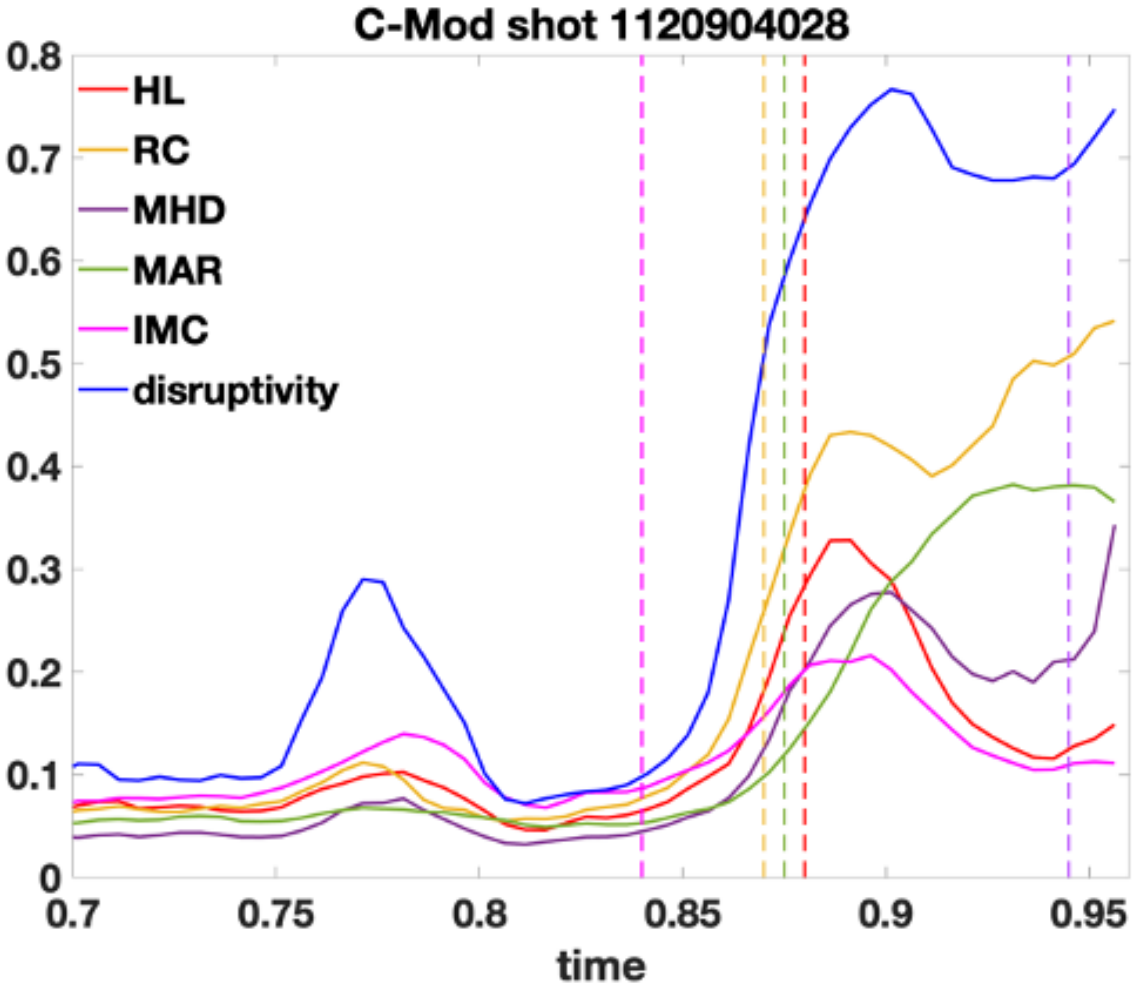

Figure 5-11: The output of the DIII-D-trained integrated model applied to a manually labeled C-Mod test shot. The onset time of all unstable events, with color corresponding to each event, are marked as vertical dashed lines in the plot. Notice that we only show the predicted levels of those events that actually happened during the flattop, and we find the predicted levels of other events are almost constant during the flattop.

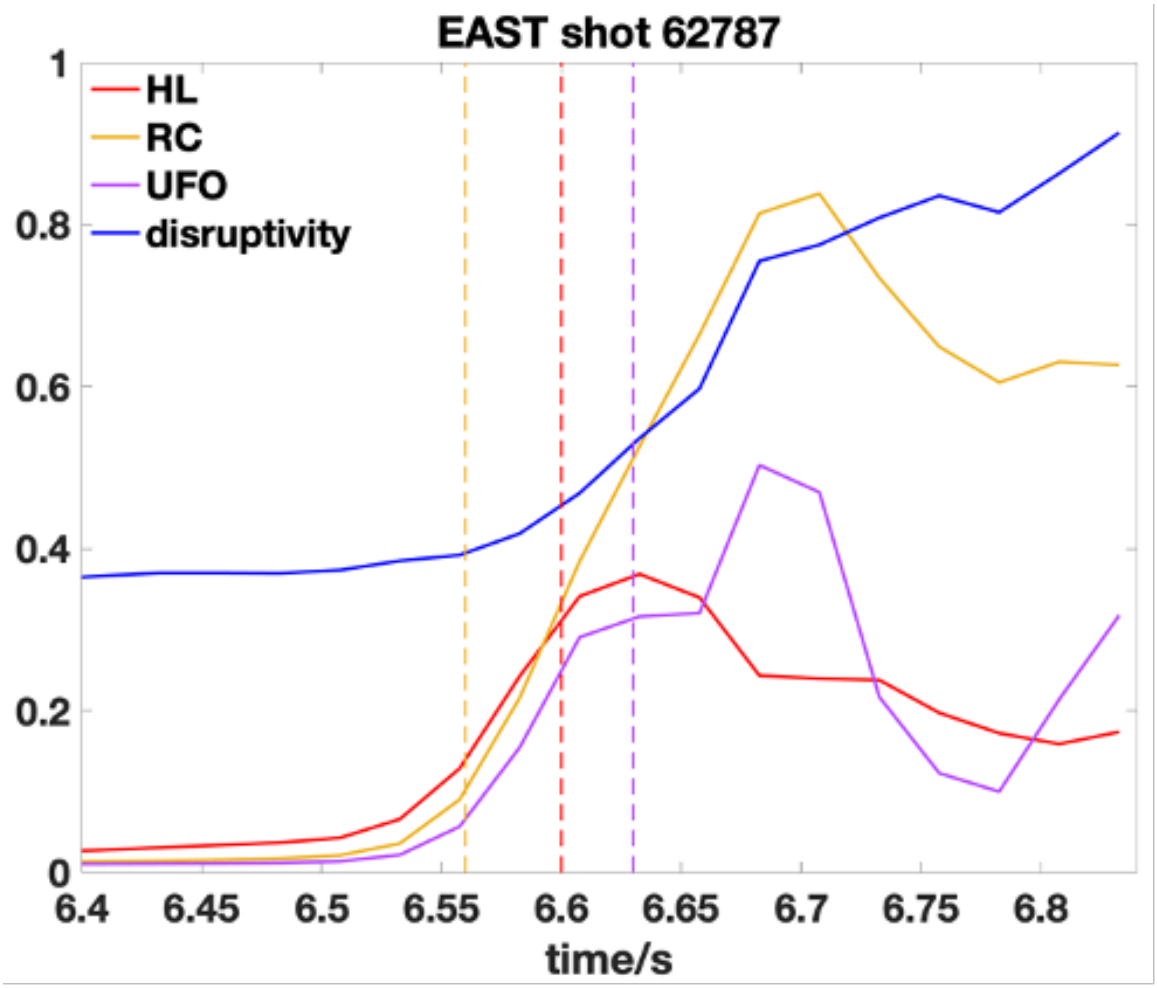

Figure 5-12: The output of the DIII-D-trained integrated model applied to a manually labeled EAST test shot. The onset time of all unstable events, with color corresponding to each event, are marked as vertical dashed lines in the plot. Notice that we only show the predicted levels of those events that actually happened during the flattop, and we find the predicted levels of other events are almost constant during the flattop.

from the machine learning perspective. Furthermore, the physics of disruption precursors has large similarity among different tokamaks, and the unstable event information provides general knowledge about disruptions. Therefore, integrated models trained with additional disruption precursor information provide better disruption prediction performance, as well as better cross-machine transferability, compared with baseline disruption predictors. Given a labeled dataset, this strategy of combining disruption predictor and disruption precursor identifier to a single integrated framework can be easily applied to up-grade any neural-network based disruption predictor and improve its performance.

5.6 Summary and future plans

In this chapter, we have discussed an iterative labeling method to automatically assign event labels to unlabeled shots, using a deep learning based event identifier and a manually labeled DIII-D database with a few hundred disruptive shots. Given the fact that all manually labeled shots are disruptive, while we need both disruptive and non-disruptive shots to train the integrated DL model, we have used the iterative labeling method to construct a training database with 160 manually labeled disruptive DIII-D shots and 900 automatically labeled non-disruptive DIII-D shots (with generated event labels). The generated event label might be biased, because the HDL-EI model used to generate these event labels is trained using only 160 manually labeled disruptive shots, all from the 2015-2016 DIII-D campaigns. In this context, we assume that the limited in-formation available on the statistical representation of DIII-D disruption dynamics might lead to a biased dataset. However, given the manually labeled disruptive set available together with the automatically generated labeled non-disruptive discharge set,, we have developed an integrated deep learning framework that can output the disruptivity score and unstable event levels simultaneously. Through numerical experiments, the integrated model is found to give higher disruption prediction accuracy, as well as longer warning time, compared with the baseline version aimed solely at predicting disruptions. The cross-machine numerical studies using C-Mod, DIII-D, and EAST data further demonstrate that the integrated model can provide better cross-machine transferability, and the integrated model, trained using data from one device, can qualitatively identify disruption precursors on a different tokamak. All of these conclusions confirm the close correlation between disruption prediction and disruption precursor identification tasks, and suggest that the physics mechanism of disruption related events shares large similarity across different tokamaks. Therefore, combining a disruption predictor and disruption precursor identifier into a single model is a promising strategy for the development of disruption predictors for future devices, and it highlights the importance of including

unstable event information when we construct the database for data-driven disruption prediction studies.

References - Chapter 5

[1] K. J. Montes, C. Rea, R. S. Granetz, R. A. Tinguely, N. Eidietis, O. M. Meneghini, D. L. Chen, B. Shen, B. J. Xiao, K. Erickson, and M. D. Boyer. Machine learning for disruption warnings on Alcator C-Mod, DIII-D, and EAST. *Nuclear Fusion*, 59(9):096015, 2019.

[2] C. Rea, K. J. Montes, K. G. Erickson, R. S. Granetz, and R. A. Tinguely. A real-time machine learning-based disruption predictor in DIII-D. *Nuclear Fusion*, 59(9):096016, 2019.

[3] Yichen Fu, David Eldon, Keith Erickson, Kornee Kleijwegt, Leonard Lupin-Jimenez, Mark D. Boyer, Nick Eidietis, Nathaniel Barbour, Olivier Izacard, and Egemen Kolemen. Machine learning control for disruption and tearing mode avoidance. *Physics of Plasmas*, 27(2):022501, 2020.

[4] Jesús Vega, Sebastián Dormido-Canto, Juan M. López, Andrea Murari, Jesús M. Ramírez, Raúl Moreno, Mariano Ruiz, Diogo Alves, and Robert Felton. Results of the JET real-time disruption predictor in the ITER-like wall campaigns. *Fusion Engineering and Design*, 88(6-8):1228–1231, Oct 2013.

[5] G. A. Rattá, J. Vega, and A. Murari. Viability Assessment of a Cross-Tokamak AUG-JET Disruption Predictor. *Fusion Science and Technology*, 74(1-2):13–22, 8 2018.

[6] A. Murari, M. Lungaroni, E. Peluso, P. Gaudio, J. Vega, S. Dormido-Canto, M. Baruzzo, and M. Gelfusa. Adaptive predictors based on probabilistic SVM for real time disruption mitigation on JET. *Nuclear Fusion*, 58(5):056002, 5 2018.

[7] W. Zheng, F.R. Hu, M. Zhang, Z.Y. Chen, X.Q. Zhao, X.L. Wang, P. Shi, X.L. Zhang, X.Q. Zhang, Y.N. Zhou, Y.N. Wei, and Y. Pan. Hybrid neural network for density limit disruption prediction and avoidance on J-TEXT tokamak. *Nuclear Fusion*, 58(5):056016, May 2018.

[8] C.G Windsor, G. Pautasso, C. Tichmann, R.J Buttery, T.C Hender, and the ASDEX Upgrade Contributors, JET EFDA and Team. A cross-tokamak neural network disruption predictor for the JET and ASDEX Upgrade tokamaks. *Nuclear Fusion*, 45(5):337, May 2005.

[9] Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568:526–531, 2019.

[10] J. Zhu, C. Rea, K. J. Montes, R. Granetz, R. Sweeney, and R. A. Tinguely. Hybrid deep learning architecture for general disruption prediction across tokamaks. *Nuclear Fusion*, 61(2):026007, 2020.

[11] Fernanda G. Rimini, Diogo Alves, Gilles Arnoux, Matteo Baruzzo, Eva Belonohy, Ivo Carvalho, Robert Felton, Emmanuel Joffrin, Peter Lomas, Paul McCullen, Andre Neto, Isabel Nunes, Cedric Reux, Adam Stephen, Daniel Valcarcel, and Sven Wiesen. The development of safe high current operation in JET-ILW. *Fusion Engineering and Design*, 96-97:165–170, 2015. Proceedings of the 28th Symposium On Fusion Technology (SOFT-28).

[12] G Pautasso, C.J Fuchs, O Gruber, C.F Maggi, M Maraschek, T Pütterich, V Ro-

hde, C Wittmann, E Wolfrum, P Cierpka, M Beck, and the ASDEX Upgrade Team. Plasma shut-down with fast impurity puff on ASDEX Upgrade. *Nuclear Fusion*, 47(8):900–913, jul 2007.

[13] Cédric Reux, Michael Lehnen, Uron Kruezi, Stefan Jachmich, Peter Card, Kalle Heinola, Emmanuel Joffrin, Peter J. Lomas, Stefan Marsen, Guy Matthews, Valeria Riccardo, Fernanda Rimini, and Peter de Vries. Use of the disruption mitigation valve in closed loop for routine protection at JET. *Fusion Engineering and Design*, 88(6):1101–1104, 2013. Proceedings of the 27th Symposium On Fusion Technology (SOFT-27); Liège, Belgium, September 24-28, 2012.

[14] E. J. Strait, J. L. Barr, M. Baruzzo, J. W. Berkery, R. J. Buttery, P. C. De Vries, N. W. Eidietis, R. S. Granetz, J. M. Hanson, C. T. Holcomb, D. A. Humphreys, J. H. Kim, E. Kolemen, M. Kong, M. J. Lanctot, M. Lehnen, E. Lerche, N. C. Logan, M. Maraschek, M. Okabayashi, J. K. Park, A. Pau, G. Pautasso, F. M. Poli, C. Rea, S. A. Sabbagh, O. Sauter, E. Schuster, U. A. Sheikh, C. Sozzi, F. Turco, A. D. Turnbull, Z. R. Wang, W. P. Wehner, and L. Zeng. Progress in disruption prevention for ITER. *Nuclear Fusion*, 59(11):112012, 2019.

[15] S. A. Sabbagh, J. W. Berkery, Y. S. Park, J. H. Ahn, J. Bialek, Y. Jiang, J. D. Riquezes, J. G. Bak, S. H. Hahn, J. Kim, J. Ko, J. Lee, S. W. Yoon, C. Ham, A. Kirk, L. Kogan, D. Ryan, A. Thornton, M. Boyer, K. Erickson, Z. Wang, V. Klevarova, and G. Pautasso. Disruption Event Characterization and Forecasting in Tokamaks and Expansion to Real-Time Application*. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP17.004, Jan 2020.

[16] M. Maraschek, A. Gude, V. Igochine, H. Zohm, E. Alessi, M. Bernert, C. Cianfarani, S. Coda, B. Duval, B. Esposito, S. Fietz, M. Fontana, C. Galperti, L. Giannone, T. Goodman, G. Granucci, L. Marelli, S. Novak, R. Paccagnella, G. Pautasso, P. Piovesan, L. Porte, S. Potzel, C. Rapson, M. Reich, O. Sauter, U. Sheikh, C. Sozzi, G. Spizzo, J. Stober, and W. Treutterer. Path-oriented early reaction to approaching disruptions in ASDEX Upgrade and TCV in view of the future needs for ITER and DEMO. *Plasma Physics and Controlled Fusion*, 60(1):14047, 2018.

[17] U.A. Sheikh, B.P. Duval, C. Galperti, M. Maraschek, O. Sauter, C. Sozzi, G. Granucci, M. Kong, B. Labit, A. Merle, N. Rispoli, and and. Disruption avoidance through the prevention of NTM destabilization in TCV. *Nuclear Fusion*, 58(10):106026, Aug 2018.

[18] A. Piccione, J. W. Berkery, S. A. Sabbagh, and Y. Andreopoulos. Physics-guided machine learning approaches to predict the ideal stability properties of fusion plasmas. *Nuclear Fusion*, 60(4):046033, 2020.

[19] K.J. Montes, C. Rea, R.A. Tinguely, R. Sweeney, J. Zhu, and R.S. Granetz. A semi-supervised machine learning detector for physics events in tokamak discharges. *Nuclear Fusion*, 61(2):026022, Jan 2021.

[20] C. Rea, R. S. Granetz, K. Montes, R. A. Tinguely, N. Eidietis, J. M. Hanson, and B. Sammuli. Disruption prediction investigations using Machine Learning tools on DIII-D and Alcator C-Mod. *Plasma Physics and Controlled Fusion*, 60(8):084004, 8 2018.

[21] A. Pau, A. Fanni, B. Cannas, S. Carcangiu, G. Pisano, G. Sias, P. Sparapani, M. Baruzzo, A. Murari, F. Rimini, M. Tsalas, and P. C. de Vries. A First Analysis of JET Plasma Profile-Based Indicators for Disruption Prediction and Avoidance. *IEEE Transactions on Plasma Science*, 46(7):2691–2698, 7 2018.

[22] Cristina Rea and Robert S. Granetz. Exploratory Machine Learning Studies for Disruption Prediction Using Large Databases on DIII-D. *Fusion Science and Technology*, 74(1-2):89–100, 8 2018.

[23] P. C. De Vries, M. F. Johnson, B. Alper, P. Buratti, T. C. Hender, H. R. Koslowski, and V. Riccardo. Survey of disruption causes at JET. *Nuclear Fusion*, 51(5):053018, 2011.

[24] Jinxiang Zhu, Cristina Rea, RS Granetz, ES Marmar, KJ Montes, Ryan Sweeney, RA Tinguely, DL Chen, Biao Shen, BJ Xiao, et al. Scenario adaptive disruption prediction study for next generation burning-plasma tokamaks. *Nuclear Fusion*, 61(11):114005, 2021.

[25] Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks*, pages 195–201. Springer, 1995.

[26] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[27] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3):121–136, 1975.

[28] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[29] Simon Haykin. Neural networks: A comprehensive foundation, 3rd edn. 1999.

[30] MP Perrone and LN Cooper. ª when networks disagree: Ensemble methods for hybrid neural networks, º neural networks for speech and image processing. *Chapman-Hall*, 1993.

[31] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[32] E. M. Hollmann, P. B. Aleynikov, T. Fülöp, D. A. Humphreys, V. A. Izzo, M. Lehnen, V. E. Lukash, G. Papp, G. Pautasso, F. Saint-Laurent, and J. A. Snipes. Status of research toward the ITER disruption mitigation system. *Physics of Plasmas*, 22(2):021802, 2 2015.

[33] S.P. Gerhardt, D.S. Darrow, R.E. Bell, B.P. LeBlanc, J.E. Menard, D. Mueller, A.L. Roquemore, S.A. Sabbagh, and H. Yuh. Detection of disruptions in the high- β spherical torus NSTX. *Nuclear Fusion*, 53(6):063021, 6 2013.

[34] P. C. de Vries, G. Pautasso, D. Humphreys, M. Lehnen, S. Maruyama, J. A. Snipes, A. Vergara, and L. Zabeo. Requirements for triggering the ITER disruption mitigation system. *Fusion Science and Technology*, 69(2):471–484, 2016.

Chapter 6

Empirical boundary detection of $n = 1$ tearing mode onset for DIII-D

The locked n=1 tearing mode (TM) is one of the key precursors that can lead to disruptions, and the ability to predict $n = 1$ TMs is highly desirable for ITER and SPARC. This is supported by the observation that the ITER baseline discharges on DIII-D are often unstable to $m/n = 2/1$ TMs that quickly lock and cause loss of confinement [1]. For disruption avoidance purpose, we want to predict the onset of the TMs since often it is too late to predict it after it developed. In response to this need for TM prediction, an empirical boundary for the $n = 1$ tearing mode (TM) is presented in this chapter. This boundary is developed via data-driven methods and verified on thousands of DIII-D discharges. It is assumed to be a linear function of plasma equilibrium parameters, including collisionality, poloidal beta, the MHD risk factor (a combination of the normalized electron temperature profile width, q_{95} and plasma elongation κ). The boundary returns with a value related to the probability of having the TM onset within 200 ms and it yeilds a shot-by-shot accuracy of about 85% in offline analysis of DIII-D data. Preliminary cross-machine analysis of TM onset prediction shows the potential applicability of the empirical boundary to C-Mod and EAST data as well, but the relative importance of the individual parameters is different for different devices. This suggests the existence of different trigger mechanisms for the TMs, implying that the boundary could be generalized using data from different tokamaks, representing different trigger mechanisms, to improve extrapolability. Finally, this newly formulated metric for proximity to $n = 1$ TM onset has been incorporated into the DIII-D real-time plasma control system (PCS), and results from real-time experiments will be discussed in Section 6.5.

6.1 Motivation of our empirical $n = 1$ TM boundary

As described in Section 1.5.5, the current driven $(2, 1)$ tearing mode is one of the most relevant instabilities leading to tokamak disruptions [2]. Furthermore, The $m/n = 2/1$ is the most dangerous island expected to show up in ITER and if it forms and locks then the plasma loses the confinement [3]. In practice, the ITER baseline scenario discharges [4] on DIII-D are often unstable to $m/n = 2/1$ TMs that quickly lock and cause complete loss of confinement [1]. Furthermore, from the manually labelled DIII-D disruption database [5], that includes all disruptive shots the 2015-2016 campaigns, roughly 77% of disruptive shots have an $n = 1$ locked mode before final thermal quench, and roughly 61% of disruptive shots have $n = 1$ TMs before the final thermal quench (see Table 5.2 for details). To date, we have studied TM prediction through two main approaches: data-driven versus physics-driven (or model-based). On one hand, there are many existing theories that describe the seed, onset and growth of TMs [6–9]. Models based on these theories have been incorporated into the DECAF algorithm, and they have shown good performance for NSTX-U and KSTAR [10, 11]. However, none of these theories can give a complete understanding of the n=1 TM. Furthermore, since these theories are usually related to some plasma parameters like Δ' that are related to the second derivative of current density and hence are hard to accurately obtain in real-time, the n=1 prediction models based on these theories are hard to implement in PCS. On the other hand, given the availability of a large amount of experimental data from decades of tokamak operation, data-driven models can be good candidates for developing a TM predictor. For example, a tree ensemble tearing model predictor has been developed for DIII-D, and has provided $> 80\%$ accuracy in an offline tests [12]. However, these data-driven methods usually don't have a closed solution, which makes interpretation difficult from the physics perspective. In turn, the limited physics interpretability renders extrapolability of these methods uncertain.

In this chapter, an empirical boundary for $n = 1$ TM on DIII-D prediction is developed via a data-driven workflow. Two datasets used in the development of TM boundary, scenario agnostic (SA) and ITER baseline scenario (IBS) $n = 1$ onset databases, will be introduced in Section 6.2. Then, in Section 6.4, the data-driven workflow and two important model selection techniques will be discussed. The SA and IBS $n = 1$ TM boundaries are presented are presented in Section 2.2.1. The offline accuracy of these two boundaries are evaluated using the corresponding test sets. Moreover, a preliminary cross-machine $n = 1$ TM prediction study is discussed in Section 2.2.1 also. Finally, the results of a dedicated TM avoidance experiment on DIII-D, using IBS TM boundary and the Off Normal Fault Response (ONFR) system [13], are presented in Section 2.2.

Table 6.1: The $n = 1$ onset dataset composition

Dataset	Num unstable shots	Num stable shots
SA training	1214	1264
SA testing	502	617
IBS training	100	90
IBS testing	66	60

6.2 The $n = 1$ TM onset databases

Our empirical $n = 1$ TM boundaries are developed using two $n = 1$ TM onset databases, the SA database and the IBS database. The IBS discharges on DIII-D comprise a series of DIII-D experiments in which the ITER shape was scaled to DIII-D dimensions, while operating with similar normalized current and β_N as in the IBS [14]. These IBS discharges usually have torque ($\approx -0.2\,\mathrm{N\,m}\sim4.2\,\mathrm{N\,m}$), $I_N \equiv \frac{I_p}{aB_{tor}} = 1.41$, $\beta_N \approx 1.75 \sim 2.25$ and $q_{95} \approx 3.1$. For the SA $n = 1$ database, we scan the **n1rms** signal (n = 1 rms amplitude of perturbed magntic field measured from toroidal array of Mirnov probes, a proxy $n = 1$ rotating mode amplitude) for the flattop of all DIII-D discharges from the 2015 to 2018 campaigns. We label a discharge as $n = 1$ unstable if the **n1rms** signal exceeds 10 Gauss for at least 25 ms during the flattop, and label a discharge as $n = 1$ stable if the **n1rms** signal is smaller than 4 Gauss throughout the whole flattop. The $n = 1$ onset time for each unstable discharge is set to the first flattop time slice that has **n1rms** signal greater than 10 Gauss. For the IBS $n = 1$ database, we manually label the $n = 1$ stability and $n = 1$ onset times for all IBS discharges from the 2011 to 2021 DIII-D campaigns. The SA and IBS databases are then divided into training sets and test sets. The dataset compositions are summarized in Table 6.1. For each unstable shot in the training set, we assign a label of **1** (close to TM onset) to all time slices that are within 200 ms of the $n = 1$ onset. This label assignment scheme fits in a supervised classification framework. For each stable shot in the training set, we assign a label of **0** to all flattop time slices. The 200 ms time threshold here is chosen based on suggestions obtained from our communication with tokamak operators that are familiar with TM on DIII-D. In addition, 200 ms is consistent with DIII-D confinement time (≈150 ms).

The 14 plasma signals included in the analysis are summarized in Table 6.2. Of these 14 signals, 10 are dimensionless, and they can be directly compared among the different tokamaks. Since our goal is to construct an interpretable symbolic boundary, we want to find the most relevant features from the above parameters, to improve the model's ability to be generalized, and to improve its physics interpretability.

Table 6.2: Plasma signals considered in the data-driven $n = 1$ onset studies [15]

Signal description	Symbol
$\frac{\textbf{Plasma current}-\textbf{programmed plasma current}}{\textbf{Programmed plasma current}}$	`ip-error-fraction`
$\frac{\textbf{Electron density}}{\textbf{Greenwald density}}$	n_G
Plasma elongation	κ
Poloidal beta	β_p
$\frac{\textbf{Radiated power}}{\textbf{Input power}}$	`radiated-fraction`
Normalized internal inductance	`li`
Safety factor at 95% flux surface	`q95`
Safety factor at magnetic axis	`q0`
Fitted half width of the T_e profile from Thomson scattering normalized by minor radius	`Te-width-norm`
Dimensionless collisionality	ν_*
Loop Voltage amplitude (V)	`v-loop`
Plasma current amplitude (MA)[a]	I_p
Toroidal magnetic field amplitude (T)[a]	B_{tor}
n = 2 rotating MHD mode (Gauss)	`n2rms`

[a]We only consider the absolute value of plasma current and toroidal magnetic field. The first 10 signals are dimensionless signals and all remaining signals are dimensional plasma signals

6.3 The data-driven workflow for $n = 1$ TM boundary discovery

Our data-driven workflow for $n = 1$ TM boundary discovery has three stages. In the first step, we investigate nonlinear combinations of all available 14 features that have high predictive power for the n=1 TM onset. In the second step, we add multiple nonlinear combinations to the original feature list, and fit a baseline logistic regression model (LR) to all features. In the third step, we use statistical model selection techniques to simplify the baseline model and get the final symbolic boundary. Two key model selection techniques, used in both the first and third steps, are backward feature elimination and probabilistic model selection. In this section, we will first introduce these two model selection techniques and then describe the details of the data-driven process.

6.3.1 Backward feature elimination

The backward feature elimination technique is an iterative process; the key metric used in this process is P-value hypothesis testing [16]. The process starts by testing the predictive power of each feature under a selected fitting of model criterion. For this step, all features are included. Then, one by one, it removes the variables that make the smallest contributions to the predictive model. Elimination is repeated until all remaining features have P-value greater than a preset significance level. Below are the steps used to execute the backward elimination:

1. Set a significance level (e.g. 0.05 which is a common practice)
2. Fit an LR model to the initial feature list.
3. Eliminate the least important feature.
    - Do the hypothesis test for each feature, and calculate the corresponding P-values.
    - Identify the feature with lowest P-value and check whether its P-value is above the significance level. If yes, remove the identified feature from the list.
4. Retrain the LR model using the updated feature list, and repeat step 3.
5. Terminate the whole process when the P-values of all features are below the significance level, and return the trained LR model.

Removing irrelevant features via the backward feature elimination technique can significantly speed up the training time, reduce model's complexity and improve model's generalization ability and interpretability. However, it also has some drawbacks. Firstly, the backward elimination technique does not consider any co-dependency between two input features. Secondly, after removing a feature from the list during the backward elimination process, that feature cannot be selected again. Since we usually apply backward feature elimination to a initial model with many reluctant features, the benefits of applying this technique will greatly outweigh its drawbacks. The co-dependency of features should be considered before applying this technique.

6.3.2 Probabilistic model selection

Probabilistic model selection is an analytical technique for scoring and choosing a model based on both its performance on the training dataset and the complexity of the model. Model accuracy is usually evaluated using a probabilistic metric, such as negative log-likelihood under the maximum likelihood estimation (MLE). Model complexity is usually evaluated using the number of parameters of the model (degrees of freedom). Although a major limitation of probabilistic model selection is that it does not take the uncertainty of the results into account and it tends, in practice, to choose overly simplified models, a clear advantage of it is that we don't need a separate test set, i.e. all of the data can be used to fit the model.

There are three main statistical metrics used in probabilistic model selection for evaluating the model's performance and complexity. These three metrics are the Akaike Information Criterion (AIC) [17], the Bayesian Information Criterion (BIC) [18] and the Minimum Description Length (MDL) [19]. It can be shown that each of these three metrics isproportional to the others, making them effectively equivalent. In this study, we will use BIC, which is derived from Bayesian probability theory and it works for models fitted under the MLE framework. The BIC metric is calculated for logistic regression as follows:

$$\mathbf{BIC} = -2 * \mathbf{LL} + \mathbf{k} * \log(\mathbf{N}) \tag{6.1}$$

where **LL** is the log-likelihood of the model on the training set, **N** is the number of samples in the training set, and **k** is the number of free parameters in the model. We want to minimize this metric and the model that has the lowest BIC will be selected. Note that, given a family of models, including the true model, the probability that minimizing the BIC metric will yield the true model approaches 1 as the number of training samples **N** goes to infinity [20].

6.3.3 The three stages of the data-driven workflow

The first stage of our data-driven workflow is tailored to find nonlinear combinations of the 14 original features that have high predictive capability. To do this, we first set a BIC score threshold. Then, to fit a nonlinear combination of our features using the linear LR model, we can take the logarithm of each feature, and normalize these logarithms by their means and standard deviations. After that, we fit an LR model to these logarithms, prune the model using backward feature elimination and then calculate the BIC score of the identified nonlinear feature combination. If the BIC score is higher than our preset threshold, we have successfully found the first nonlinear combination, and we remove all features included in that combination from the feature list. If the BIC score is lower than the preset threshold, we consider the first stage concluded and return all identified nonlinear feature combinations. We repeat this process until all combinations are explored, after that we consider the first stage concluded.

As the second stage of our workflow, we first add all identified nonlinear combinations to the original feature list. Then we fit an LR model using this combined feature list. This is our baseline model. In stage three, we apply the backward feature elimination and probabilistic model selection techniques to the baseline model, finally yielding the simplified symbolic TM boundary.

6.4 The symbolic $n = 1$ TM boundaries

6.4.1 The SA $n = 1$ TM boundary

By applying the data-driven workflow, mentioned in Section 6.4, to the DIII-D SA dataset, we get a symbolic $n = 1$ TM boundary that works for DIII-D SA shots. The resulting formula for this SA $n = 1$ TM boundary is:

$$\textbf{TM risk} = \beta_p - 0.35\nu_* - 0.09\frac{\sqrt{\texttt{q95}*\texttt{Te-width-norm}}}{\kappa} + 2.03\frac{I_p}{B_{tor}} - 2.07 \quad (6.2)$$

To evaluate the performance of the above SA $n = 1$ TM boundary, we want to test it on the DIII-D SA test set. To this aim, a shot-by-shot testing scheme is developed to simulate alarms triggered in the DIII-D PCS. For a given a test shot, we can calculate the TM risk for each time slice during the discharge flattop. If the flattop TM risk exceeds a preset threshold for at least **T** ms, the test shot is classified as TM unstable, and the warning time is recorded for truly $n = 1$ unstable shots, defined as the difference between the alarm time and the actual $n = 1$ onset time (t_{onset}). A successfully detected $n = 1$ unstable DIII-D SA test shot is shown in Figure 6-1. If

the alarm time is greater than 0 for an $n = 1$ unstable test shot, this shot is classified as a true positive. If the flattop TM risk of an $n = 1$ stable test shot does not exceed the preset threshold for at least **T** ms, this shot is a true negative shot. For all of our offline analysis, we chose $\mathbf{T} = 25\,\text{ms}$. Considering our task is to classify stable vs unstable to TM time slices, our SA boundary achieves TPR= 0.88 and TNR= 0.89 on the SA test set.

To further evaluate the relation between flattop TM risk and $n = 1$ onset probability, we calculate the maximum TM risk level for each shot in the DIII-D SA set. The maximum TM risk level for each shot is defined as the maximum value of TM threshold that the flattop TM risk exceeds for at least **T** ms. Given the maximum TM risk level for each shot, we can divide the SA shots according to their maximum TM risk level, and calculate the fraction of the unstable shots in each bin. The relation between maximum TM risk level and empirical $n = 1$ onset probability of SA shots is shown in Figure 6-2. From this figure, we find the empirical $n = 1$ onset probability goes to 1 when the maximum TM risk level goes above roughly 0.5, while the empirical $n = 1$ onset probability goes below 0.05 when the maximum TM risk level goes below roughly (-0.5).

6.4.2 The IBS $n = 1$ TM boundary

If we substitute typical IBS parameters into our SA boundary, we find the resulting TM levels lie in the range [-0.05, 0.1]. From Figure 6-2, this range corresponds to TM probability ≈ 0.5 and it agrees with the observed TM frequency in IBS shots, which suggests that a scenario agnostic boundary doesn't work in IBS; we need a new boundary, fitted with IBS specific data. To establish a symbolic TM boundary for IBS, we applied our data-driven workflow to the IBS training set. The resulting formula for our IBS $n = 1$ TM boundary is:

$$\textbf{TM risk} = 0.20\beta_p + 0.01\nu_* + 1.81\frac{\sqrt{\texttt{q95}*\texttt{Te-width-norm}}}{\kappa} + 1.69\frac{I_p}{B_{tor}} - 3.1 \quad (6.3)$$

Notice that this boundary only works for IBS and the coefficients of IBS TM boundary are clearly different from SA TM boundary which implies the fact that the prevalence TM mechanism in IBS is different from typical TM mechanism in other scenarios. To evaluate the performance of the above IBS $n = 1$ TM boundary, we next test it on the DIII-D IBS test set, using the same shot-by-shot testing scheme described in Section 6.4.1. A comparison between a detected $n = 1$ unstable DIII-D IBS shot, and a true negative DIII-D IBS shot, is shown in Figure 6-3. It is clear that the TM risk of the unstable IBS shot shown in Figure 6-3 exceeds the threshold for roughly 200 ms before the final $n = 1$ onset, while the stable IBS shot shown in Figure 6-3

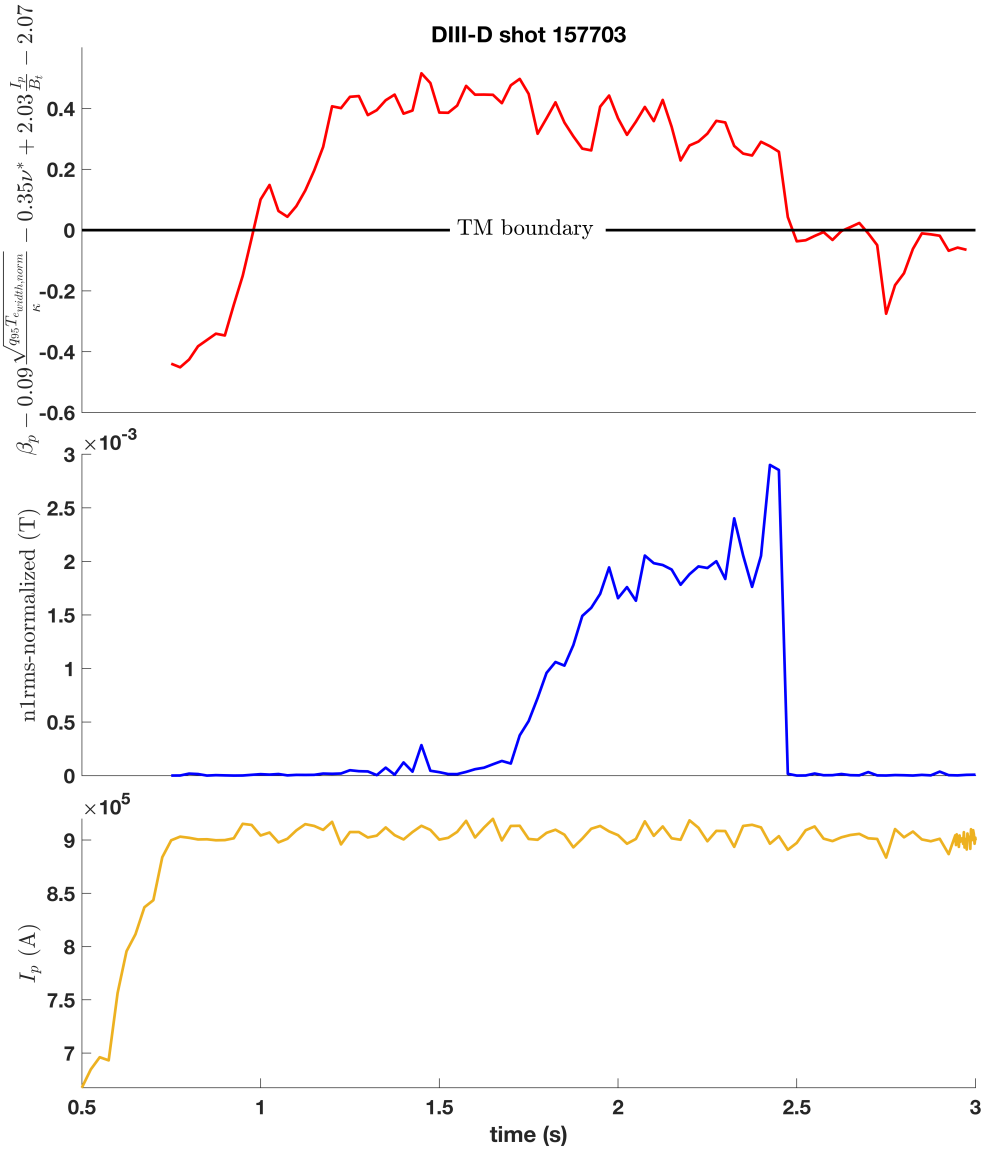

Figure 6-1: An example of a true positive shot from DIII-D SA test set. The upper panel shows the time trace of the calculated TM risk and the preset TM threshold. The middle panel shows the time trace of the $n = 1$ rotating MHD mode proxy (normalized n1rms signal). The lower panel shows the time trace of the plasma current I_p. The TM risk first exceeds the threshold at around 500 ms before the actual $n = 1$ onset, and it goes back below the threshold just before $n = 1$ rotating mode locks.

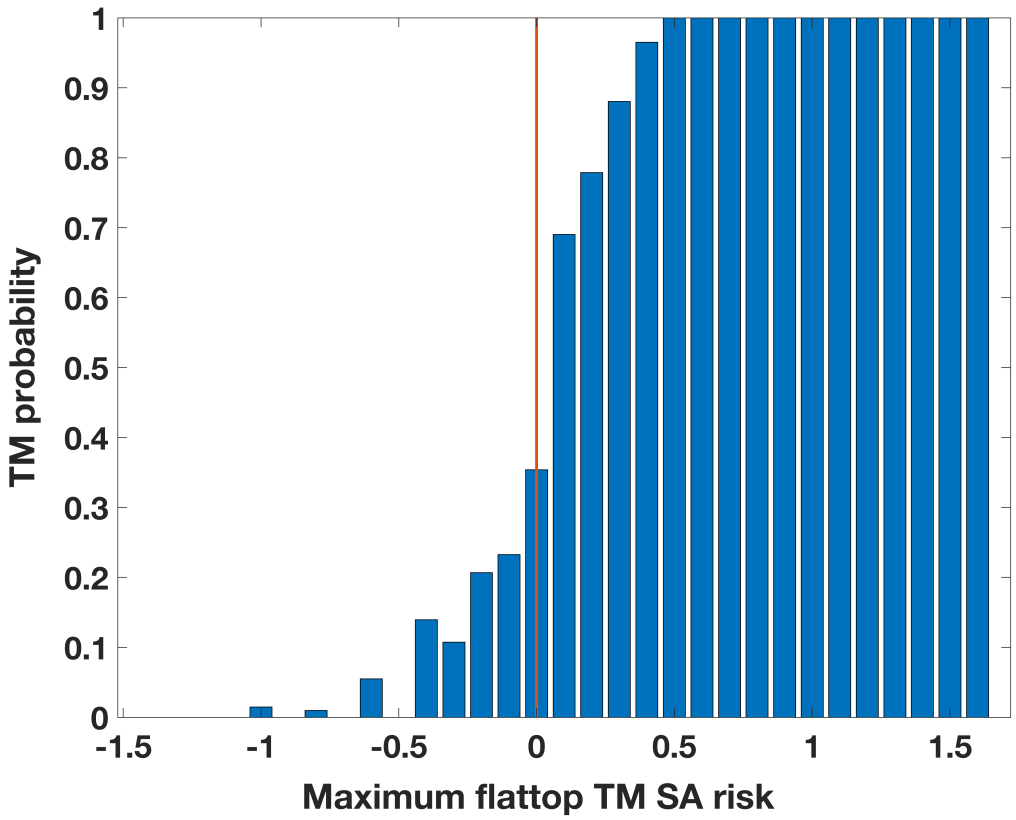

Figure 6-2: The empirical TM probability vs. maximum flattop TM risk level for all DIII-D SA shots.

never continuously exceeds the threshold for $\mathbf{T} = 25\,\text{ms}$. Our IBS boundary analysis achieves TPR= 0.70 and TNR= 0.72 on the IBS test set, which is worse than the performance of the SA TM boundary on SA test set. The worsening of prediction performance with respect to IBS comes from the fact that IBS is more unstable to TM compared to most of other DIII-D scenarios. Indeed, we find the TM risk difference between TM unstable and TM stable shots is very small in IBS, which means the margin between TM stable and TM unstable regime is very narrow. This observation agrees with the fact that predicting TM in IBS is significantly more difficult than in the SA case.

6.4.3 Preliminary cross-machine $n = 1$ TM boundary study

To test the cross-machine ability of our data-driven workflow, we apply our TM boundary trained on DIII-D data to other tokamaks. Establishing a cross-machine TM boundary requires us only to consider those plasma parameters that are comparable among different devices, namely the dimensionless parameters. Therefore, we remove all dimensional parameters (the last 4 features in Table 6.2) from the original 14, and apply our data-driven workflow to the DIII-D SA dataset with the 10 remaining dimensionless features. The resulting cross-machine $n = 1$ TM boundary formula is:

$$\textbf{TM risk} = \beta_p - 1.14\nu_* - 1.36\frac{\sqrt{\texttt{q95 * Te-width-norm}}}{\kappa} + 0.51 \qquad (6.4)$$

Again, we find that the coefficients of cross-machine TM boundary are clearly different from both SA TM boundary and IBS TM boundary. Considering the fact that the cross-machine TM boundary is also trained on SA DIII-D data with only dimensionless features, the large difference in coefficient suggests the large impact of original feature list to final obtained boundary and it might implies the power of dimensional features for TM prediction on a chosen device. We applied this boundary to the EAST disruption warning database described in Section 3.2. The obtained flattop TM risk distribution, from more than 10000 EAST shots, is shown in Figure 6-4. As we can see from this plot, the flattop TM risk distribution from EAST is even more negative than flattop TM risk distribution for DIII-D stable shots, and it has a very pronounced negative tail. This result suggests that the flattop TM risk rarely exceeds 0 for EAST, and it implies that $n = 1$ TM onset is much less frequent on EAST than on DIII-D. This agrees with our observation that EAST discharges rarely have natural $n = 1$ rotating modes. A similar analysis of C-Mod flattop data shows that C-Mod also has a more negative TM risk relative to DIII-D,also agreeing with the observation that $n = 1$ modes are much less frequently seen in C-Mod discharges

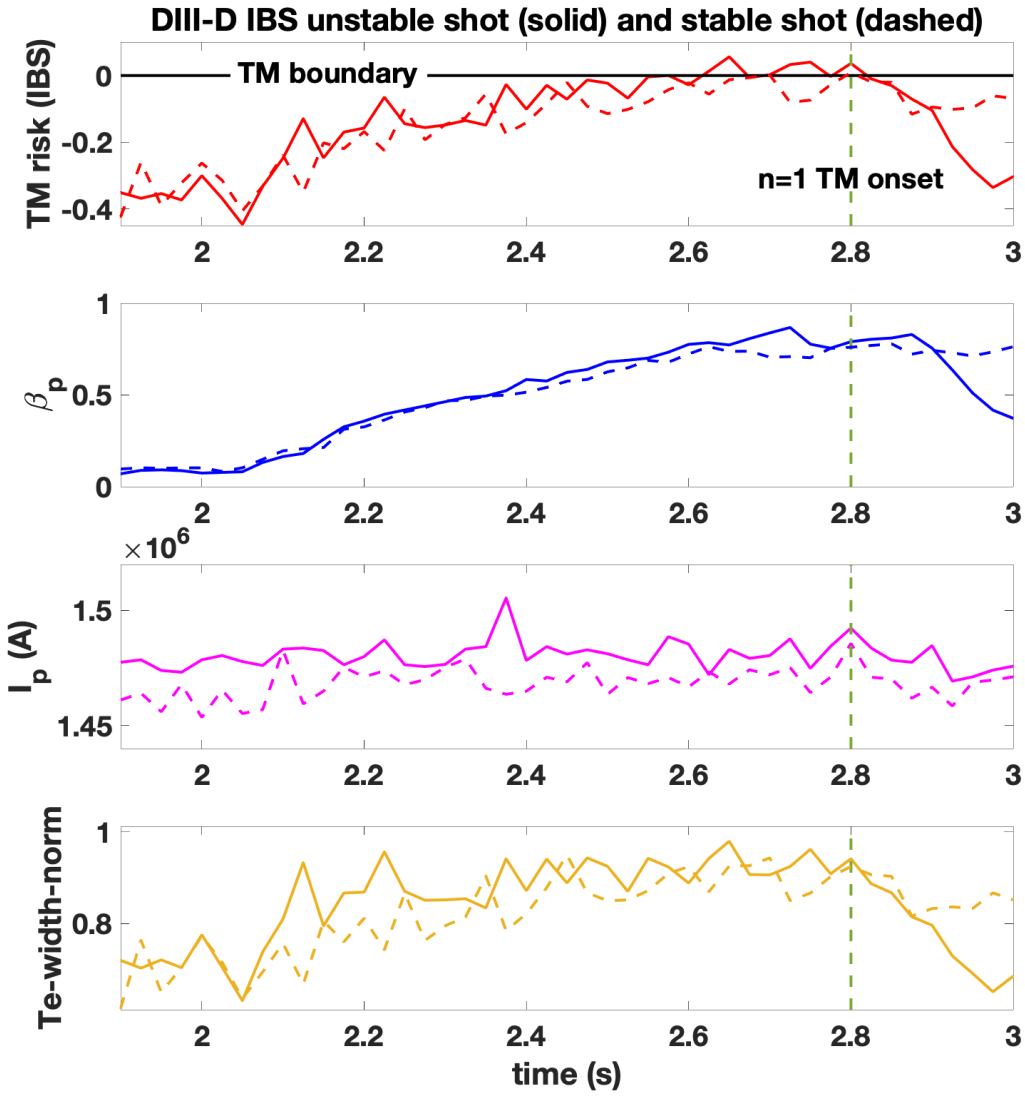

Figure 6-3: A comparison between a detected $n = 1$ unstable DIII-D IBS shot (solid line) and a true negative DIII-D IBS shot (dashed line). The first panel shows the time trace of the TM risk, calculated using IBS TM boundary, and the corresponding TM risk boundary. The remaining panel shows the time traces of other key parameters included in the IBS TM boundary formula. As can be seen, the TM risk of the unstable shot exceeds the threshold for roughly 200 ms before final $n = 1$ onset, while the stable IBS shot never continuously exceeds the threshold for $\mathbf{T} = 25$ ms. The differences between stable and unstable shots mainly come from the difference of the `Te-width-norm` time traces.

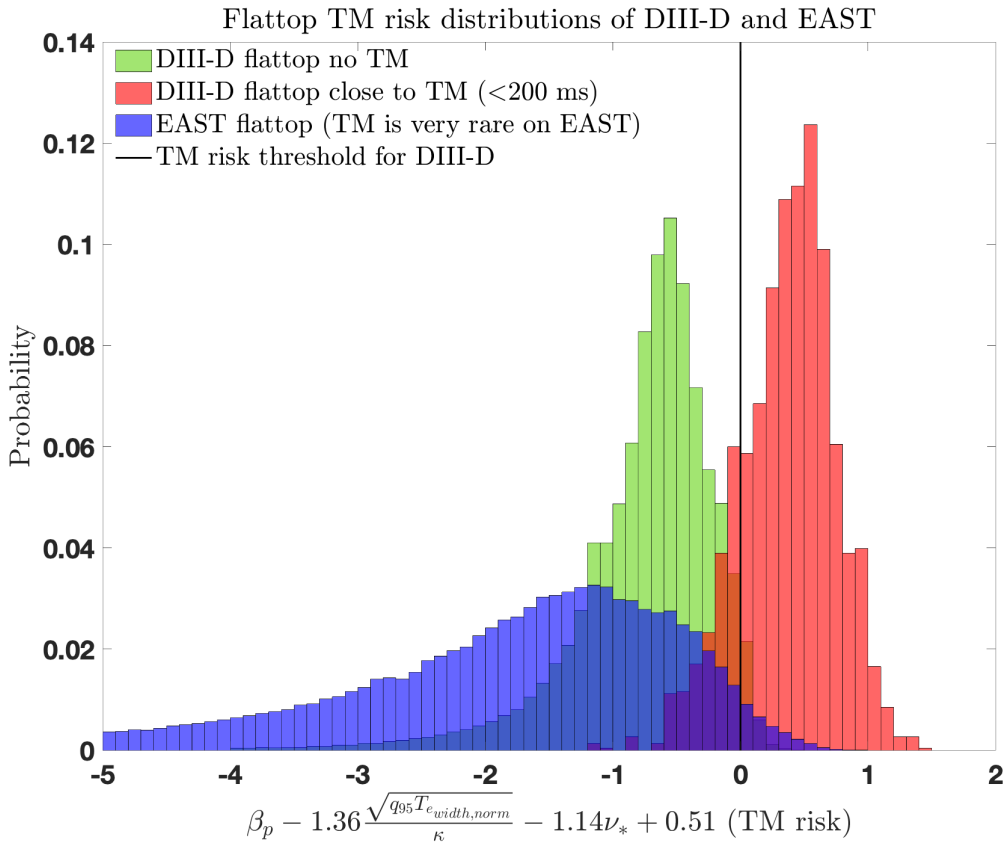

Figure 6-4: The empirical flattop TM risk distributions for the EAST disruption warning database (more than 10000 shots, blue), DIII-D $n = 1$ stable shots ($\approx$ 2000 shots, green) and DIII-D $n = 1$ unstable shots ($\approx$ 2000 shots, red). From this plot, it is clear that the flattop TM risk distribution for EAST is even more negative than flattop TM risk distribution of DIII-D stable shots. This agrees with the observation that $n = 1$ TM onset is much less frequent in EAST than in DIII-D

than in DIII-D.

6.4.4 Summary of symbolic $n = 1$ TM boundaries

In this section, we applied our data-driven workflow for $n = 1$ TM boundary discovery to DIII-D SA, DIII-D IBS and DIII-D dimensionless datasets to obtain the SA, IBS and cross-machine $n = 1$ TM boundary. The offline analysis of these three boundaries suggests these boundaries achieve reasonably good performance on the corresponding test set, and it proves the effectiveness of our data-driven workflow. In addition, we find the three TM boundaries have very different weights for different parameters (e.g. the weight of $\nu*$ in the SA boundary is much larger than it is in the IBS boundary). This observation suggests the existence of multiple TM trigger mechanisms, implying that the boundary could better be generalized using data from different tokamaks

representing different trigger mechanisms to improve its extrapolability.

6.5 Real-time $n = 1$ TM avoidance experiments in IBS

Prediction and avoidance of $n = 1$ TM onset in IBS is a challenging task on DIII-D. To tackle this problem, our IBS $n = 1$ TM onset boundary has been incorporated into the real-time DIII-D PCS, and dedicated DIII-D experiments were conducted in July 2022 to demonstrate the real-time applicability of TM onset prediction and TM avoidance through the integration of this stability metric with the Off Normal Fault Response (ONFR) [13]. ONFR is a robust supervisory system implemented on DIII-D for comprehensive disruption avoidance and machine protection. The key idea of our experiments was to sweep some of the parameters for the TM risk, both during and between shots, and scan thresholds in TM risk and time delay ($\Delta\mathbf{T}$) to trigger the ONFR, aiming for TM avoidance or to initiate a soft landing of the shot. Since ν_* shows a large correlation with other plasma parameters (q_{95}, β_p), and it has a relatively small impact on the tearing mode levels, our basic experimental approach is to change β_p and q_{95} between shots. To do this, we scanned B_{tor} or and I_p and keep $\beta_N = 2.5$ constant. By increasing I_p or decreasing B_{tor} during the discharge, we can trigger the $n = 1$ TM onset, and then identify the proper TM risk threshold and trigger time delay for triggering the ONFR to avoiding TM onset. A successful TM avoidance experiment is shown in Figure 6-5. In this experiment, we ramped up I_p from 1.35MA to 1.5MA to trigger an $n = 1$ TM onset. For the first shot, the ONFR was disabled; the TM risk exceeded the threshold at $\approx 300\,\text{ms}$ before $n = 1$ TM onset. For the second shot, we repeated the previous shot, but with the ONFR enabled. The ONFR is triggered at $\approx 2.1\,\text{s}$ resulting an early I_p ramp-down (in line with the "soft landing" idea) and the $n = 1$ TM onset is successfully avoided. Our experiments suggest the real-time IBS TM boundary achieves similar accuracy compared with the offline tests described in Section 6.4.3, making it a powerful tool for TM avoidance in DIII-D IBS discharges. By collecting more IBS data, and retraining the IBS TM boundary, we should be able to continuously improve the accuracy of our TM boundary.

6.6 Summary and future plans

In this chapter, a data-driven workflow for symbolic $n = 1$ TM boundary detection is discussed in detail. This data-driven workflow is then applied to DIII-D SA and DIII-D IBS regimes, to derive corresponding $n = 1$ TM boundaries. These boundaries

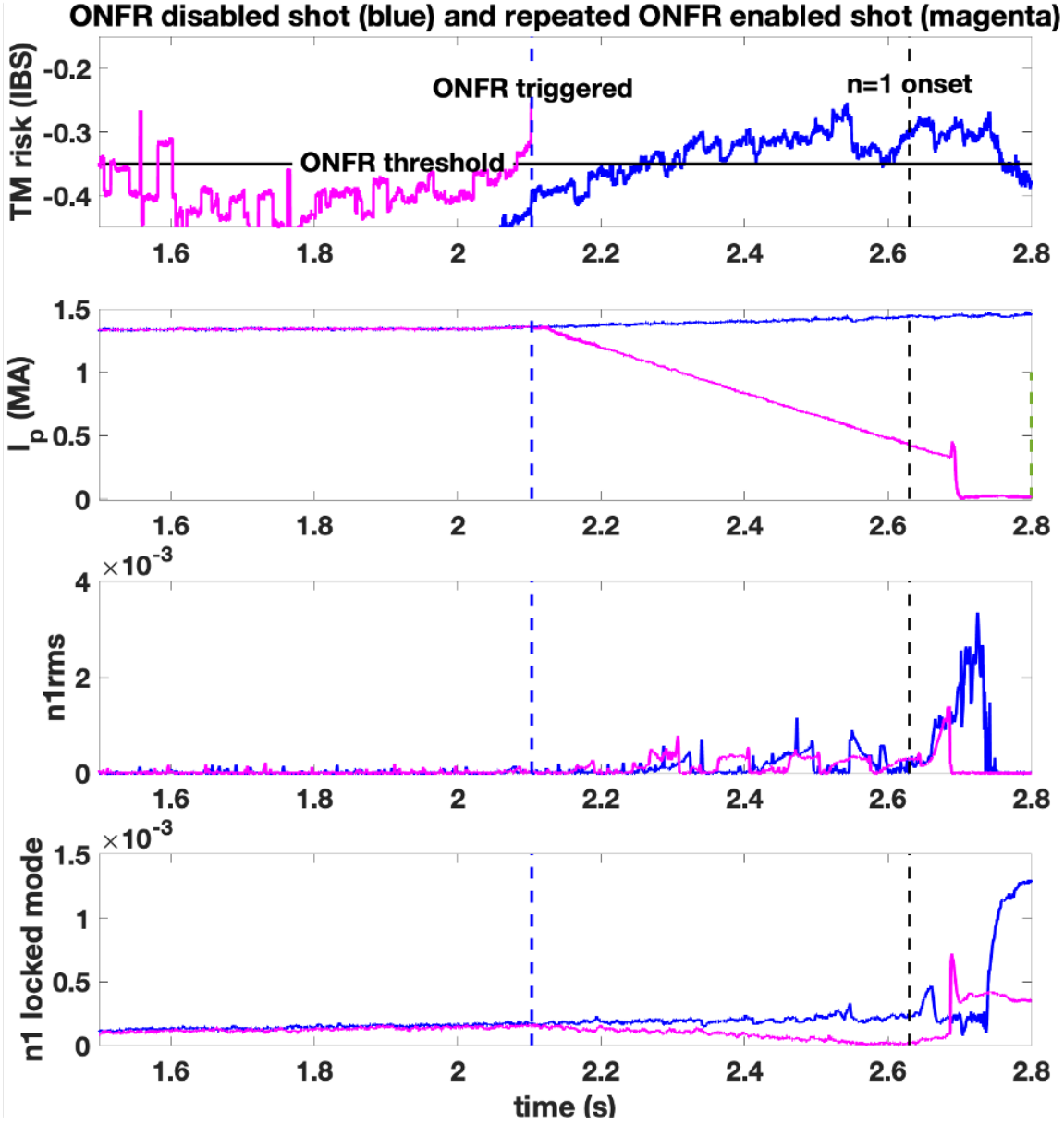

Figure 6-5: A successful TM avoidance experiment is shown in this figure. In this experiment, we ramp up I_p from 1.35MA to 1.5MA to trigger $n = 1$ TM onset. From the top to the bottom panel, we show the time traces of TM risk, I_p, n1rms and $n = 1$ locked mode signal for an ONFR disabled shot (blue) and a repeated shot with ONFR enabled (magenta). For the ONFR disabled shot, the TM risk exceeded the boundary $\approx 300\,\text{ms}$ before the $n = 1$ TM onset. For the repeated, ONFR enabled shot, the ONFR is triggered at $\approx 2.1\,\text{s}$. The system switches the plasma control system to start an early I_p rampdown, and the $n = 1$ TM onset is successfully avoided.

achieves good shot-by-shot accuracy in offline tests. The preliminary cross machine study using our data-driven workflow, and the DIII-D dimensionless dataset, shows the promising cross machine potential of our method. In addition, a real-time TM instability metric for IBS is integrated into the DIII-D PCS. Dedicated DIII-D IBS experiments using this real-time metric, combined with real-time response from the ONFR, demonstrate the real-time TM avoidance capability of our IBS TM boundary approach.

References - Chapter 6

[1] GL Jackson, TC Luce, WM Solomon, F Turco, RJ Buttery, AW Hyatt, JS DeGrassie, EJ Doyle, JR Ferron, RJ La Haye, et al. Long-pulse stability limits of the iter baseline scenario. *Nuclear Fusion*, 55(2):023004, 2015.

[2] P.H. Rutherford. Tearing modes in tokamaks. In B. Coppi, G.G. Leotta, D. Pfirsch, R. Pozzoli, and E. Sindoni, editors, *Physics of Plasmas Close to Thermonuclear Conditions*, pages 129–142. Pergamon, 1981.

[3] RJ La Haye, C Chrystal, EJ Strait, JD Callen, CC Hegna, EC Howell, M Okabayashi, and RS Wilcox. Disruptive neoclassical tearing mode seeding in diii-d with implications for iter. *Nuclear Fusion*, 62(5):056017, 2022.

[4] E.J. Doyle, J.C. DeBoo, J.R. Ferron, G.L. Jackson, T.C. Luce, M. Murakami, T.H. Osborne, J.-M Park, P.A. Politzer, H. Reimerdes, Robert Budny, Thomas Casper, C.D. Challis, R.J. Groebner, C.T. Holcomb, A.W. Hyatt, R.J. Haye, G.R. McKee, T.W. Petrie, and Lunwu Zeng. Demonstration of iter operational scenarios on diii-d. *Nuclear Fusion*, 50:075005, 06 2010.

[5] K.J. Montes, C. Rea, R.A. Tinguely, R. Sweeney, J. Zhu, and R.S. Granetz. A semi-supervised machine learning detector for physics events in tokamak discharges. *Nuclear Fusion*, 61(2):026022, Jan 2021.

[6] Paul H. Rutherford. Nonlinear growth of the tearing mode. *Physics of Fluids*, 16:1903–1908, 1973.

[7] Roscoe B. White, D. A. Monticello, Marshall N. Rosenbluth, and B. V. Waddell. Saturation of the tearing mode. *The Physics of Fluids*, 20(5):800–805, 1977.

[8] R. J. La Haye, S. Günter, D. A. Humphreys, J. Lohr, T. C. Luce, M. E. Maraschek, C. C. Petty, R. Prater, J. T. Scoville, and E. J. Strait. Control of neoclassical tearing modes in diii–d. *Physics of Plasmas*, 9(5):2051–2060, 2002.

[9] R. J. La Haye. Neoclassical tearing modes and their control. *Physics of Plasmas*, 13(5):055501, 2006.

[10] J. D. Riquezes, S. A. Sabbagh, J. W. Berkery, Y. S. Park, J. H. Ahn, Y. Jiang, J. Butt, E. Fredrickson, and J. G. Bak. Rotating MHD Mode Analysis Including Real-time data on KSTAR Supporting Disruption Event Characterization and Forecasting. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP17.007, Jan 2020.

[11] Yousung Park, Steve Sabbagh, Jaeheon Ahn, B Park, Hyun-Seok Kim, John Berkery, Jim Bialek, Yanzheng Jiang, Jun-Gyo BAK, Alan Glasser, Jisung Kang, Jaehyun Lee, Hyunsun Han, Sang-Hee Hahn, Youngmu Jeon, B. Jung, Hyeon Kyun Park, Zhirui Wang, Jong-Kyu Park, and S Yoon. Analysis of mhd stability and active mode control on kstar for high confinement, disruption-free plasma. *Nuclear Fusion*, 60, 02 2020.

[12] Yichen Fu, David Eldon, Keith Erickson, Kornee Kleijwegt, Leonard Lupin-Jimenez, Mark D. Boyer, Nick Eidietis, Nathaniel Barbour, Olivier Izacard, and Egemen Kolemen. Machine learning control for disruption and tearing mode avoidance. *Physics of Plasmas*, 27(2):022501, 2020.

[13] N.W. Eidietis, W. Choi, S.H. Hahn, D.A. Humphreys, B.S. Sammuli, and M.L. Walker. Implementing a finite-state off-normal and fault response system for

disruption avoidance in tokamaks. *Nuclear Fusion*, 58(5):056023, mar 2018.

[14] F. Turco, T.C. Luce, W. Solomon, G. Jackson, G.A. Navratil, and J.M. Hanson. The causes of the disruptive tearing instabilities of the ITER baseline scenario in DIII-D. *Nuclear Fusion*, 58(10):106043, Sep 2018.

[15] J. Zhu, C. Rea, K. J. Montes, R. Granetz, R. Sweeney, and R. A. Tinguely. Hybrid deep learning architecture for general disruption prediction across tokamaks. *Nuclear Fusion*, 61(2):026007, 2020.

[16] Karl Pearson F.R.S. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[17] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[18] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[19] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[20] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, second edition, 2009.

Chapter 7

Conclusions and Future Work

7.1 Summary and main contributions

This thesis focuses on the data-driven solution of major disruption and plasma instabilities prediction, a topic with implications for the success of future burning-plasma tokamaks, ITER and SPARC, and the pilot plant reactors that may follow them. The main contributions of this thesis work are the following:

- The author contributed to the **development of a hybrid deep-learning (HDL) model for cross-machine disruption prediction** [1], described in chapter 3. This included an unsupervised clustering study that highlights the advantage of the sequence-based model and it also included several qualitative conclusions that can provide guidelines for the development of data-driven disruption predictor on future devices.

- The author demonstrated a **scenario adaptive development strategy of data-driven disruption predictor for future tokamaks** [2], as discussed in chapter 4. The scenario adaptive disruption prediction studies on C-Mod, DIII-D and EAST show that data-driven predictors trained only on LP discharges perform poorly for the subsequent HP regime of the same tokamak. Dedicated cross-machine studies further suggest that matching operational parameters among devices can greatly improve prediction accuracy for the target device, which highlights the importance of building comprehensive databases that consist of different kinds of disruptive burning-plasma baseline scenario discharges from current devices. A successful strategy for data-driven disruption prediction on future tokamaks like ITER is to combine ITER LP data with HP ITER baseline discharges from other devices to train a predictor with enough accuracy to help ITER conserve its disruption budget during the early stage of its HP operation; as ITER's HP operation continues, add HP ITER

data to the training set. Retraining the predictor using this combined dataset will boost the predictor performance towards ITER's long-term requirements [3].

- The author developed an **integrated deep learning framework for unstable event identification and disruption prediction of tokamak plasmas**, as discussed in chapter 5. The numerical disruption prediction studies demonstrate that the integrated model gives higher disruption prediction accuracy, as well as longer warning time, compared with the baseline version aimed solely at predicting disruptions, and it also has better cross-machine transferability. The integrated model, trained using data from one device, can qualitatively identify disruption precursors on a different tokamak. These results suggest that combining a disruption predictor and disruption precursor identifier into a single model is a promising strategy for the development of disruption predictors on future devices, and it highlights the importance of including unstable event information when we construct the database for data-driven disruption prediction studies.

- The author developed a **data-driven workflow for symbolic $n = 1$ TM boundary discovery**, as discussed in chapter 6. The DIII-D SA and IBS TM boundaries, that show good offline accuracy, are obtained by applying this workflow to DIII-D SA and DIII-D IBS datasets. The preliminary cross machine study using this data-driven workflow and DIII-D dimensionless dataset shows the good cross machine potential of our method. Moreover, a real-time algorithm based on the IBS TM boundary is integrated into DIII-D PCS. Dedicated DIII-D IBS experiments using this algorithm and ONFR demonstrate the real-time TM avoidance capability of the IBS TM boundary.

In addition to the above contributions of this thesis, the author's research has also supported related projects in disruption prediction, mitigation, and avoidance. This has resulted in multiple co-authored articles, including one validating the SPARC physics basis [4, 5] and a second on semi-supervised learning detector for unstable events in tokamak discharges [6].

7.2 Future efforts

Research efforts to follow up on the results of this thesis can focus in several directions. Firstly, the author wants to extend the deep-learning based disruption prediction studies (HDL model and integrated DL model) to data from additional tokamaks, including JET, AUG, JT-60SA and KSTAR. Including more data from

various tokamaks should further test the cross-machine applicability of the disruption predictor, and the model's cross-machine ability can be continually increased by retraining these model using data from more devices. In addition, the author wants to extend scenario adaptive study to these additional tokamaks and investigate whether the findings reported in **??** hold for the ITER baseline discharges on additional existing tokamaks. Secondly, given a more powerful GPU cluster, the author plans to fine-tune the hyper-parameters of the deep-learning disruption predictor to further boost the performance of the model. Thirdly, given the iterative labeling method described Section 5.3, and considering the value of unstable event information, the author plans to manually label unstable events for few hundreds DIII-D, EAST and C-Mod disruptive and non-disruptive shots. Then the author can further expand the manually labeled databases by using iterative labeling method with the original manually labeled datasets, to automatically label more shots. The expanded DIII-D, C-Mod and EAST databases will allow the author to investigate the disruption prediction capability limit of integrated model on existing devices. In addition, cross-machine numerical experiments using these databases and integrated model can further confirm the efficiency of the labeling method, and also allow the author to develop a more robust cross-machine integrated model using databases from all three tokamaks. Fourthly, a real-time test of the integrated model capabilities should be conducted on an existing tokamak (DIII-D is a good candidate) to further investigate robust disruption prediction and avoidance strategies, and to facilitate development of a disruption handling system on future tokamaks. Fifthly, the author wants to investigate the most dangerous paths to disruptions on future tokamaks, including ITER and SPARC, and perform tokamak discharge simulations for these conditions (e.g. MHD simulation). Then the author can try to apply the integrated model to the synthetic signal and quantify the performance of the predictor. The integrated DL model can be used together with the tokamak discharge simulator to find a good operating scenario (stable and high fusion power/fusion gain plasma) for future tokamaks. Sixthly, given that the IBS TM boundary has already been integrated into DIII-D PCS and it will continuously run in background during standard operations, the author wants to collect more data to estimate the aging effect of the IBS TM boundary. Seventhly, the author wants to include more predictive features like rotation shear and current well information [7] to further boost the performance of our IBS TM boundary. Eighthly, the author wants to build multi-machine TM database and develop more robust cross-machine TM boundary using this dataset. Dedicated experiments can be conducted on both DIII-D and EAST to validate the accuracy of cross-machine TM boundary. Finally, the author plans to leverage existing code basis and achievements to develop a robust data-driven disruption predictor for SPARC. The scenario adaptive strategy and qualitative cross-machine disruption prediction

conclusions presented in this thesis will facilitate the development of SPARC data-driven disruption predictor.

References - Chapter 7

[1] J. Zhu, C. Rea, K. J. Montes, R. Granetz, R. Sweeney, and R. A. Tinguely. Hybrid deep learning architecture for general disruption prediction across tokamaks. *Nuclear Fusion*, 61(2):026007, 2020.

[2] Jinxiang Zhu, Cristina Rea, RS Granetz, ES Marmar, KJ Montes, Ryan Sweeney, RA Tinguely, DL Chen, Biao Shen, BJ Xiao, et al. Scenario adaptive disruption prediction study for next generation burning-plasma tokamaks. *Nuclear Fusion*, 61(11):114005, 2021.

[3] P. C. de Vries, G. Pautasso, D. Humphreys, M. Lehnen, S. Maruyama, J. A. Snipes, A. Vergara, and L. Zabeo. Requirements for triggering the ITER disruption mitigation system. *Fusion Science and Technology*, 69(2):471–484, 2016.

[4] A. J. Creely and M. J. et al. Greenwald. Overview of the SPARC tokamak. *Journal of Plasma Physics*, 86(5):865860502, 2020.

[5] R. Sweeney, A. J. Creely, J. Doody, T. Fülöp, D. T. Garnier, R. Granetz, M. Greenwald, L. Hesslow, J. Irby, V. A. Izzo, R.J. La Haye, N.C. Logan, K. Montes, C. Paz-Soldan, C. Rea, R.A. Tinguely, O. Vallhagen, and J. Zhu. MHD stability and disruptions in the SPARC tokamak. *Journal of Plasma Physics*, 86(5):865860507, 2020.

[6] K.J. Montes, C. Rea, R.A. Tinguely, R. Sweeney, J. Zhu, and R.S. Granetz. A semi-supervised machine learning detector for physics events in tokamak discharges. *Nuclear Fusion*, 61(2):026022, Jan 2021.

[7] F. Turco, T.C. Luce, W. Solomon, G. Jackson, G.A. Navratil, and J.M. Hanson. The causes of the disruptive tearing instabilities of the ITER baseline scenario in DIII-D. *Nuclear Fusion*, 58(10):106043, Sep 2018.

List of Figures

List of Tables

References

[1] IEA. World Energy Outlook 2020. Technical report, International Energy Agency, 2020.

[2] BP. Statistical Review of World Energy. Technical Report 69, British Petroleum, 2020.

[3] JGJ Olivier and JAHW Peters. Trends in global co2 and total greenhouse gas emissions. *PBL Netherlands Environmental Assessment Agency: The Hague, The Netherlands*, 2020.

[4] J. Freidberg. *Plasma Physics and Fusion Energy*. Cambridge University Press, 2007.

[5] J D Lawson. Some criteria for a power producing thermonuclear reactor. *Proceedings of the Physical Society. Section B*, 70(1):6–10, jan 1957.

[6] John Wesson. *Tokamaks; 4th ed.* International series of monographs on physics. Oxford Univ. Press, Oxford, 2011.

[7] VP Smirnov. Tokamak foundation in ussr/russia 1950–1990. *Nuclear fusion*, 50(1):014003, 2009.

[8] V Mukhovatov, M Shimada, A N Chudnovskiy, A E Costley, Y Gribov, G Federici, O Kardaun, A S Kukushkin, A Polevoi, V D Pustovitov, Y Shimomura, T Sugie, M Sugihara, and G Vayakis. Overview of physics basis for ITER. *Plasma Physics and Controlled Fusion*, 45(12A):A235–A252, Nov 2003.

[9] A. J. Creely and M. J. et al. Greenwald. Overview of the SPARC tokamak. *Journal of Plasma Physics*, 86(5):865860502, 2020.

[10] J. Freidberg. *Ideal MHD*. Cambridge University Press, 2014.

[11] Harold Grad and Hanan Rubin. Hydromagnetic equilibria and force-free fields. *Journal of Nuclear Energy (1954)*, 7(3-4):284–285, 1958.

[12] V. D. Shafranov. Plasma Equilibrium in a Magnetic Field. *Reviews of Plasma Physics*, 2:103, Jan 1966.

[13] M. Murakami, J.D. Callen, and L.A. Berry. Some observations on maximum densities in tokamak experiments. *Nuclear Fusion*, 16(2):347–348, apr 1976.

[14] Martin Greenwald, J. L. Terry, S. M. Wolfe, S. Ejima, M.G. Bell, S. M. Kaye, and G. H. Neilson. A new look at density limits in tokamaks. *Nuclear Fusion*, 28(12):2199–2207, 1988.

[15] P. C. De Vries, M. F. Johnson, B. Alper, P. Buratti, T. C. Hender, H. R. Koslowski, and V. Riccardo. Survey of disruption causes at JET. *Nuclear Fusion*, 51(5):053018, 2011.

[16] V. P. Lipschultz, R. La Bombard, P. Marmar, R. Pickrell, R. R. Terry, R. R. Watterson, and S. M. Wolfe. Marfe: An edge plasma phenomenon. *Nuclear Fusion*, 24(8):977–988,

1984.

[17] Q. Teng, D.P. Brennan, L. Delgado-Aparicio, D.A. Gates, J. Swerdlow, and R.B. White. A predictive model for the tokamak density limit. *Nuclear Fusion*, 56(10):106001, Jul 2016.

[18] A. Sykes, M. F. Turner, and S. Patel. Proceedings of the 11th european conference on controlled fusion and plasma physics. volume 2, page 363. European Physical Society.

[19] F. Troyon and R. Gruber. A semi-empirical scaling law for the β-limit in tokamaks. *Physics Letters A*, 110(1):29–34, 1985.

[20] J. E. Menard, M. G. Bell, R. E. Bell, D. A. Gates, S. M. Kaye, B. P. LeBlanc, R. Maingi, S. A. Sabbagh, V. Soukhanovskii, and D. Stutman. Aspect ratio scaling of ideal no-wall stability limits in high bootstrap fraction tokamak plasmas. *Physics of Plasmas*, 11(2):639–646, 2004.

[21] P. Piovesan, J. M. Hanson, P. Martin, G. A. Navratil, F. Turco, J. Bialek, N. M. Ferraro, R. J. La Haye, M. J. Lanctot, M. Okabayashi, C. Paz-Soldan, E. J. Strait, A. D. Turnbull, P. Zanca, M. Baruzzo, T. Bolzonella, A. W. Hyatt, G. L. Jackson, L. Marrelli, L. Piron, and D. Shiraki. Tokamak operation with safety factor $q_{95} < 2$ via control of MHD stability. *Phys. Rev. Lett.*, 113:045003, Jul 2014.

[22] M. Okabayashi and G. Sheffield. Vertical stability of elongated tokamaks. *Nuclear Fusion*, 14(2):263–265, apr 1974.

[23] E.A. Lazarus, J.B. Lister, and G.H. Neilson. Control of the vertical instability in tokamaks. *Nuclear Fusion*, 30(1):111–141, jan 1990.

[24] V Riccardo, P Noll, and SP Walker. Forces between plasma, vessel and tf coils during avdes at jet. *Nuclear fusion*, 40(10):1805, 2000.

[25] R.S Granetz, I.H Hutchinson, J Sorci, J.H Irby, B LaBombard, and D Gwinn. Disruptions and halo currents in Alcator C-Mod. *Nuclear Fusion*, 36(5):545–556, May 1996.

[26] P.H. Rutherford. Tearing modes in tokamaks. In B. Coppi, G.G. Leotta, D. Pfirsch, R. Pozzoli, and E. Sindoni, editors, *Physics of Plasmas Close to Thermonuclear Conditions*, pages 129–142. Pergamon, 1981.

[27] R. Sweeney, W. Choi, R.J. La Haye, S. Mao, K.E.J. Olofsson, and F.A. Volpe. Statistical analysis of m / n = 2/1 locked and quasi-stationary modes with rotating precursors at DIII-D. *Nuclear Fusion*, 57(1):016019, Jan 2017.

[28] S.P. Gerhardt, D.S. Darrow, R.E. Bell, B.P. LeBlanc, J.E. Menard, D. Mueller, A.L. Roquemore, S.A. Sabbagh, and H. Yuh. Detection of disruptions in the high- β spherical torus NSTX. *Nuclear Fusion*, 53(6):063021, 6 2013.

[29] M.F.F. Nave and J.A. Wesson. Mode locking in tokamaks. *Nuclear Fusion*, 30(12):2575–2583, dec 1990.

[30] R Fitzpatrick. Interaction of tearing modes with external structures in cylindrical geometry (plasma). *Nuclear Fusion*, 33(7):1049–1084, jul 1993.

[31] Alberto Loarte. Effects of divertor geometry on tokamak plasmas. *Plasma Physics and Controlled Fusion*, 43(6):R183–R224, may 2001.

[32] S K Rathgeber, R Fischer, S Fietz, J Hobirk, A Kallenbach, H Meister, T Pütterich, F Ryter, G Tardini, and E Wolfrum and. Estimation of profiles of the effective ion charge at ASDEX upgrade with integrated data analysis. *Plasma Physics and Controlled Fusion*, 52(9):095008, aug 2010.

[33] M B Chowdhuri, R Manchanda, J Ghosh, K A Jadeja, Kaushal M Patel, Vinay Kumar, Ketan M Patel, P K Atrey, Y Shankara Joisa, S B Bhatt, and R L Tanna

and. Investigation of the behavior of effective charge of ADITYA tokamak plasmas. *Plasma Physics and Controlled Fusion*, 62(3):035015, Feb 2020.

[34] J. Zhu, C. Rea, K. J. Montes, R. Granetz, R. Sweeney, and R. A. Tinguely. Hybrid deep learning architecture for general disruption prediction across tokamaks. *Nuclear Fusion*, 61(2):026007, 2020.

[35] *Fusion Physics*. Non-serial Publications. INTERNATIONAL ATOMIC ENERGY AGENCY, Vienna, 2012.

[36] V. Riccardo and A. Loarte. Timescale and magnitude of plasma thermal energy loss before and during disruptions in JET. *Nuclear Fusion*, 45(11):1427–1438, 2005.

[37] Jochen Linke, Juan Du, Thorsten Loewenhoff, Gerald Pintsuk, Benjamin Spilker, Isabel Steudel, and Marius Wirtz. Challenges for plasma-facing components in nuclear fusion. *Matter and Radiation at Extremes*, 4(5):056201, 2019.

[38] S.N. Gerasimov, P. Abreu, M. Baruzzo, V. Drozdov, A. Dvornova, J. Havlicek, T.C. Hender, O. Hronova, U. Kruezi, X. Li, T. Markovič, R. Pánek, G. Rubinacci, M. Tsalas, S. Ventre, F. Villone, and L.E. Zakharov and. JET and COMPASS asymmetrical disruptions. *Nuclear Fusion*, 55(11):113006, sep 2015.

[39] C. E. Myers, N. W. Eidietis, S. N. Gerasimov, S. P. Gerhardt, R. S. Granetz, T. C. Hender, and G. Pautasso. A multi-machine scaling of halo current rotation. *Nuclear Fusion*, 58(1), 2018.

[40] Boris N. Breizman, Pavel Aleynikov, Eric M. Hollmann, and Michael Lehnen. Physics of runaway electrons in tokamaks. *Nuclear Fusion*, 59(8):083001, jun 2019.

[41] M.N Rosenbluth and S.V Putvinski. Theory for avalanche of runaway electrons in tokamaks. *Nuclear Fusion*, 37(10):1355–1362, Oct 1997.

[42] Allen H. Boozer. Theory of tokamak disruptions. *Physics of Plasmas*, 19(5), 2012.

[43] B. Bazylev, G. Arnoux, W. Fundamenski, Yu. Igitkhanov, and M. Lehnen. Modeling of runaway electron beams for JET and ITER. *Journal of Nuclear Materials*, 415(1, Supplement):S841–S844, 2011. Proceedings of the 19th International Conference on Plasma-Surface Interactions in Controlled Fusion.

[44] C. Reux, V. Plyusnin, B. Alper, D. Alves, B. Bazylev, E. Belonohy, A. Boboc, S. Brezinsek, I. Coffey, J. Decker, P. Drewelow, S. Devaux, P.C. de Vries, A. Fil, S. Gerasimov, L. Giacomelli, S. Jachmich, E.M. Khilkevitch, V. Kiptily, R. Koslowski, U. Kruezi, M. Lehnen, I. Lupelli, P.J. Lomas, A. Manzanares, A. Martin De Aguilera, G.F. Matthews, J. Mlynář, E. Nardon, E. Nilsson, C. Perez von Thun, V. Riccardo, F. Saint-Laurent, A.E. Shevelev, G. Sips, and C. Sozzi and. Runaway electron beam generation and mitigation during disruptions at JET-ILW. *Nuclear Fusion*, 55(9):093013, Aug 2015.

[45] T.C. Hender, J.C. Wesley, J. Bialek, A. Bondeson, A.H. Boozer, R.J. Buttery, A. Garofalo, T.P. Goodman, R.S. Granetz, Y. Gribov, O. Gruber, M. Gryaznevich, G. Giruzzi, S. Günter, N. Hayashi, P. Helander, C.C. Hegna, D.F. Howell, D.A. Humphreys, G.T.A. Huysmans, A.W. Hyatt, A. Isayama, S.C. Jardin, Y. Kawano, A. Kellman, C. Kessel, H.R. Koslowski, R.J. la Haye, E. Lazzaro, Y.Q. Liu, V. Lukash, J. Manickam, S. Medvedev, V. Mertens, S.V. Mirnov, Y. Nakamura, G. Navratil, M. Okabayashi, T. Ozeki, R. Paccagnella, G. Pautasso, F. Porcelli, V.D. Pustovitov, V. Riccardo, M. Sato, O. Sauter, M.J. Schaffer, M. Shimada, P. Sonato, E. J. Strait, M. Sugihara, M. Takechi, A.D. Turnbull, E. Westerhof, D.G. Whyte, R. Yoshino, and H. Zohm. Chapter 3: Mhd stability, operational limits and disruptions. *Nuclear Fusion*, 47(6):S128–S202, June 2007.

[46] M. Lehnen. Plasma disruption management in ITER. In *26th IAEA Fusion Energy Conference IAEA*, pages EX/P6–39, 2016.

[47] F. Turco, T.C. Luce, W. Solomon, G. Jackson, G.A. Navratil, and J.M. Hanson. The causes of the disruptive tearing instabilities of the ITER baseline scenario in DIII-D. *Nuclear Fusion*, 58(10):106043, Sep 2018.

[48] H Zohm, G Gantenbein, G Giruzzi, S Günter, F Leuterer, M Maraschek, J Meskat, AG Peeters, W Suttrop, D Wagner, et al. Experiments on neoclassical tearing mode stabilization by eccd in asdex upgrade. *Nuclear Fusion*, 39(5):577, 1999.

[49] D. Li, Q. Yu, Y. Ding, N. Wang, F. Hu, R. Jia, L. Peng, B. Rao, Q. Hu, H. Jin, M. Li, L. Zhu, Z. Huang, Z. Song, S. Zhou, J. Li, Y. He, Q. Zhang, W. Zhang, J. Dong, D. Han, W. Zheng, A. A. Bala, K. Yu, and Y. Liang and. Disruption prevention using rotating resonant magnetic perturbation on J-TEXT. *Nuclear Fusion*, 60(5):056022, Apr 2020.

[50] M. Lehnen, K. Aleynikova, P.B. Aleynikov, D.J. Campbell, P. Drewelow, N.W. Eidietis, Yu. Gasparyan, R.S. Granetz, Y. Gribov, N. Hartmann, E.M. Hollmann, V.A. Izzo, S. Jachmich, S.-H. Kim, M. Kočan, H.R. Koslowski, D. Kovalenko, U. Kruezi, A. Loarte, S. Maruyama, G.F. Matthews, P.B. Parks, G. Pautasso, R.A. Pitts, C. Reux, V. Riccardo, R. Roccella, J.A. Snipes, A.J. Thornton, and P.C. de Vries. Disruptions in ITER and strategies for their control and mitigation. *Journal of Nuclear Materials*, 463:39–48, 2015. PLASMA-SURFACE INTERACTIONS 21.

[51] D. Shiraki, N. Commaux, L. R. Baylor, N. W. Eidietis, E. M. Hollmann, C. J. Lasnier, and R. A. Moyer. Thermal quench mitigation and current quench control by injection of mixed species shattered pellets in DIII-D. *Physics of Plasmas*, 23(6):062516, 2016.

[52] T. C. Jernigan, L. A. Baylor, S. K. Combs, D. A. Humphreys, P. B. Parks, and J. C. Wesley. Massive gas injection systems for disruption mitigation on the DIII-D tokamak. In *21st IEEE/NPS Symposium on Fusion Engineering SOFE 05*, pages 1–3, 2005.

[53] M Lehnen, A Alonso, G Arnoux, N Baumgarten, SA Bozhenkov, S Brezinsek, M Brix, T Eich, SN Gerasimov, A Huber, et al. Disruption mitigation by massive gas injection in jet. *Nuclear fusion*, 51(12):123010, 2011.

[54] RA Tinguely, VA Izzo, DT Garnier, A Sundström, K Särkimäki, O Embréus, T Fülöp, RS Granetz, M Hoppe, I Pusztai, et al. Modeling the complete prevention of disruption-generated runaway electron beam formation with a passive 3d coil in sparc. *Nuclear Fusion*, 61(12):124003, 2021.

[55] M. Greenwald, A. Bader, S. Baek, M. Bakhtiari, H. Barnard, W. Beck, W. Bergerson, I. Bespamyatnov, P. Bonoli, D. Brower, D. Brunner, W. Burke, J. Candy, M. Churchill, I. Cziegler, A. Diallo, A. Dominguez, B. Duval, E. Edlund, P. Ennever, D. Ernst, I. Faust, C. Fiore, T. Fredian, O. Garcia, C. Gao, J. Goetz, T. Golfinopoulos, R. Granetz, O. Grulke, Z. Hartwig, S. Horne, N. Howard, A. Hubbard, J. Hughes, I. Hutchinson, J. Irby, V. Izzo, C. Kessel, B. LaBombard, C. Lau, C. Li, Y. Lin, B. Lipschultz, A. Loarte, E. Marmar, A. Mazurenko, G. McCracken, R. McDermott, O. Meneghini, D. Mikkelsen, D. Mossessian, R. Mumgaard, J. Myra, E. Nelson-Melby, R. Ochoukov, G. Olynyk, R. Parker, S. Pitcher, Y. Podpaly, M. Porkolab, M. Reinke, J. Rice, W. Rowan, A. Schmidt, S. Scott, S. Shiraiwa, J. Sierchio, N. Smick, J. A. Snipes, P. Snyder, B. Sorbom, J. Stillerman, C. Sung, Y. Takase, V. Tang, J. Terry, D. Terry, C. Theiler, A. Tronchin-James, N. Tsujii, R. Vieira, J. Walk, G. Wallace, A. White, D. Whyte, J. Wilson, S. Wolfe, G. Wright, J. Wright, S. Wukitch, and

S. Zweben. 20 years of research on the Alcator C-Mod tokamak. *Physics of Plasmas*, 21(11):110501, 2014.

[56] P. C. de Vries, G. Pautasso, D. Humphreys, M. Lehnen, S. Maruyama, J. A. Snipes, A. Vergara, and L. Zabeo. Requirements for triggering the ITER disruption mitigation system. *Fusion Science and Technology*, 69(2):471–484, 2016.

[57] E. J. Strait, J. L. Barr, M. Baruzzo, J. W. Berkery, R. J. Buttery, P. C. De Vries, N. W. Eidietis, R. S. Granetz, J. M. Hanson, C. T. Holcomb, D. A. Humphreys, J. H. Kim, E. Kolemen, M. Kong, M. J. Lanctot, M. Lehnen, E. Lerche, N. C. Logan, M. Maraschek, M. Okabayashi, J. K. Park, A. Pau, G. Pautasso, F. M. Poli, C. Rea, S. A. Sabbagh, O. Sauter, E. Schuster, U. A. Sheikh, C. Sozzi, F. Turco, A. D. Turnbull, Z. R. Wang, W. P. Wehner, and L. Zeng. Progress in disruption prevention for ITER. *Nuclear Fusion*, 59(11):112012, 2019.

[58] Fernanda G. Rimini, Diogo Alves, Gilles Arnoux, Matteo Baruzzo, Eva Belonohy, Ivo Carvalho, Robert Felton, Emmanuel Joffrin, Peter Lomas, Paul McCullen, Andre Neto, Isabel Nunes, Cedric Reux, Adam Stephen, Daniel Valcarcel, and Sven Wiesen. The development of safe high current operation in JET-ILW. *Fusion Engineering and Design*, 96-97:165–170, 2015. Proceedings of the 28th Symposium On Fusion Technology (SOFT-28).

[59] G Pautasso, C.J Fuchs, O Gruber, C.F Maggi, M Maraschek, T Pütterich, V Rohde, C Wittmann, E Wolfrum, P Cierpka, M Beck, and the ASDEX Upgrade Team. Plasma shut-down with fast impurity puff on ASDEX Upgrade. *Nuclear Fusion*, 47(8):900–913, jul 2007.

[60] Cédric Reux, Michael Lehnen, Uron Kruezi, Stefan Jachmich, Peter Card, Kalle Heinola, Emmanuel Joffrin, Peter J. Lomas, Stefan Marsen, Guy Matthews, Valeria Riccardo, Fernanda Rimini, and Peter de Vries. Use of the disruption mitigation valve in closed loop for routine protection at JET. *Fusion Engineering and Design*, 88(6):1101–1104, 2013. Proceedings of the 27th Symposium On Fusion Technology (SOFT-27); Liège, Belgium, September 24-28, 2012.

[61] Vitus Mertens, Gerhard Raupp, and Wolfgang Treutterer. Chapter 3: Plasma Control in ASDEX Upgrade. *Fusion Science and Technology*, 44(3):593–604, 2003.

[62] N.W. Eidietis, W. Choi, S.H. Hahn, D.A. Humphreys, B.S. Sammuli, and M.L. Walker. Implementing a finite-state off-normal and fault response system for disruption avoidance in tokamaks. *Nuclear Fusion*, 58(5):056023, mar 2018.

[63] N.M. Trang Vu, T.C. Blanken, F. Felici, C. Galperti, M. Kong, E. Maljaars, and O. Sauter. Tokamak-agnostic actuator management for multi-task integrated control with application to TCV and ITER. *Fusion Engineering and Design*, 147:111260, 2019.

[64] W. Treutterer, R. Cole, K. Lüddecke, G. Neu, C. Rapson, G. Raupp, D. Zasche, and T. Zehetbauer. ASDEX Upgrade discharge control system—a real-time plasma control framework. *Fusion Engineering and Design*, 89(3):146–154, 2014. Design and implementation of real-time systems for magnetic confined fusion devices.

[65] S. A. Sabbagh, J. W. Berkery, Y. S. Park, J. H. Ahn, J. Bialek, Y. Jiang, J. D. Riquezes, J. G. Bak, S. H. Hahn, J. Kim, J. Ko, J. Lee, S. W. Yoon, C. Ham, A. Kirk, L. Kogan, D. Ryan, A. Thornton, M. Boyer, K. Erickson, Z. Wang, V. Klevarova, and G. Pautasso. Disruption Event Characterization and Forecasting in Tokamaks and Expansion to Real-Time Application*. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP17.004, Jan 2020.

[66] J. W. Berkery, S. A. Sabbagh, R. E. Bell, S. P. Gerhardt, and B. P. LeBlanc. A reduced resistive wall mode kinetic stability model for disruption forecasting. *Physics of Plasmas*, 24(5):056103, 2017.

[67] J. D. Riquezes, S. A. Sabbagh, J. W. Berkery, Y. S. Park, J. H. Ahn, Y. Jiang, J. Butt, E. Fredrickson, and J. G. Bak. Rotating MHD Mode Analysis Including Real-time data on KSTAR Supporting Disruption Event Characterization and Forecasting. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP17.007, Jan 2020.

[68] J. Butt, S. A. Sabbagh, Y. S. Park, J. H. Ahn, J. W. Berkery, Y. Jiang, and J. D. Riquezes. ELM Detection Capability for Disruption Event Characterization and Forecasting. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP17.006, January 2020.

[69] Won Ha Ko, SW Yoon, WC Kim, JG Kwak, KR Park, YU Nam, SJ Wang, J Chung, BH Park, GY Park, et al. Kstar overview. *Bulletin of the American Physical Society*, 2022.

[70] M. Maraschek, A. Gude, V. Igochine, H. Zohm, E. Alessi, M. Bernert, C. Cianfarani, S. Coda, B. Duval, B. Esposito, S. Fietz, M. Fontana, C. Galperti, L. Giannone, T. Goodman, G. Granucci, L. Marelli, S. Novak, R. Paccagnella, G. Pautasso, P. Piovesan, L. Porte, S. Potzel, C. Rapson, M. Reich, O. Sauter, U. Sheikh, C. Sozzi, G. Spizzo, J. Stober, and W. Treutterer. Path-oriented early reaction to approaching disruptions in ASDEX Upgrade and TCV in view of the future needs for ITER and DEMO. *Plasma Physics and Controlled Fusion*, 60(1):14047, 2018.

[71] U.A. Sheikh, B.P. Duval, C. Galperti, M. Maraschek, O. Sauter, C. Sozzi, G. Granucci, M. Kong, B. Labit, A. Merle, N. Rispoli, and and. Disruption avoidance through the prevention of NTM destabilization in TCV. *Nuclear Fusion*, 58(10):106026, Aug 2018.

[72] D. Humphreys, A. Kupresanin, M. D. Boyer, J. Canik, C. S. Chang, E. C. Cyr, R. Granetz, J. Hittinger, E. Kolemen, E. Lawrence, V. Pascucci, A. Patra, and D. Schissel. Advancing fusion with machine learning research needs workshop report. *Journal of Fusion Energy*, 39:123–155, 2020.

[73] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 3 edition, 2006.

[74] Max Kuhn and Kjell Johnson. *Feature engineering and selection: A practical approach for predictive models*. CRC Press, 2019.

[75] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[76] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, second edition, 2009.

[77] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

[78] Akanksh Basavaraju, Jing Du, Fujie Zhou, and Jim Ji. A machine learning approach to road surface anomaly assessment using smartphone sensors. *IEEE Sensors Journal*, 20(5):2635–2647, 2020.

[79] K. J. Montes, C. Rea, R. S. Granetz, R. A. Tinguely, N. Eidietis, O. M. Meneghini, D. L. Chen, B. Shen, B. J. Xiao, K. Erickson, and M. D. Boyer. Machine learning for disruption warnings on Alcator C-Mod, DIII-D, and EAST. *Nuclear Fusion*, 59(9):096015, 2019.

[80] C. Rea, K. J. Montes, K. G. Erickson, R. S. Granetz, and R. A. Tinguely. A real-time

machine learning-based disruption predictor in DIII-D. *Nuclear Fusion*, 59(9):096016, 2019.

[81] Yichen Fu, David Eldon, Keith Erickson, Kornee Kleijwegt, Leonard Lupin-Jimenez, Mark D. Boyer, Nick Eidietis, Nathaniel Barbour, Olivier Izacard, and Egemen Kolemen. Machine learning control for disruption and tearing mode avoidance. *Physics of Plasmas*, 27(2):022501, 2020.

[82] Jesús Vega, Sebastián Dormido-Canto, Juan M. López, Andrea Murari, Jesús M. Ramírez, Raúl Moreno, Mariano Ruiz, Diogo Alves, and Robert Felton. Results of the JET real-time disruption predictor in the ITER-like wall campaigns. *Fusion Engineering and Design*, 88(6-8):1228–1231, Oct 2013.

[83] G. A. Rattá, J. Vega, and A. Murari. Viability Assessment of a Cross-Tokamak AUG-JET Disruption Predictor. *Fusion Science and Technology*, 74(1-2):13–22, 8 2018.

[84] A. Murari, M. Lungaroni, E. Peluso, P. Gaudio, J. Vega, S. Dormido-Canto, M. Baruzzo, and M. Gelfusa. Adaptive predictors based on probabilistic SVM for real time disruption mitigation on JET. *Nuclear Fusion*, 58(5):056002, 5 2018.

[85] R. Aledda, B. Cannas, A. Fanni, A. Pau, and G. Sias. Improvements in disruption prediction at ASDEX Upgrade. *Fusion Engineering and Design*, 96-97:698–702, 10 2015.

[86] W. Zheng, F.R. Hu, M. Zhang, Z.Y. Chen, X.Q. Zhao, X.L. Wang, P. Shi, X.L. Zhang, X.Q. Zhang, Y.N. Zhou, Y.N. Wei, and Y. Pan. Hybrid neural network for density limit disruption prediction and avoidance on J-TEXT tokamak. *Nuclear Fusion*, 58(5):056016, May 2018.

[87] C.G Windsor, G. Pautasso, C. Tichmann, R.J Buttery, T.C Hender, and the ASDEX Upgrade Contributors, JET EFDA and Team. A cross-tokamak neural network disruption predictor for the JET and ASDEX Upgrade tokamaks. *Nuclear Fusion*, 45(5):337, May 2005.

[88] A. Piccione, J. W. Berkery, S. A. Sabbagh, and Y. Andreopoulos. Physics-guided machine learning approaches to predict the ideal stability properties of fusion plasmas. *Nuclear Fusion*, 60(4):046033, 2020.

[89] Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568:526–531, 2019.

[90] Diogo R. Ferreira, Pedro J. Carvalho, Carlo Sozzi, Peter J. Lomas, and JET Contributors. Deep learning for the analysis of disruption precursors based on plasma tomography. *Fusion Science and Technology*, 76(8):901–911, 2020.

[91] R. M. Churchill, B. Tobias, and Y. Zhu. 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*, 27(6):062510, 2020.

[92] A. Pau, A. Fanni, S. Carcangiu, B. Cannas, G. Sias, A. Murari, and F. Rimini. A machine learning approach based on generative topographic mapping for disruption prevention and avoidance at JET. *Nuclear Fusion*, 59(10):106017, 2019.

[93] A. Murari, J. Vega, G.A. Rattá, G. Vagliasindi, M.F. Johnson, and S.H. Hong. Unbiased and non-supervised learning methods for disruption prediction at JET. *Nuclear Fusion*, 49(5):055028, 5 2009.

[94] Raffaele Aledda, Barbara Cannas, Alessandra Fanni, Giuliana Sias, and Gabriella Pautasso. Adaptive mapping of the plasma operational space of ASDEX Upgrade for disruption prediction. *International Journal of Applied Electromagnetics and Mechan-*

ics, 39(1-4):43–49, 2012.

[95] Y. Wei, J. W. Brooks, R. Chandra, J. P. Levesque, Boting Li, A. Saperstein, I. G. Stewart, M. E. Mauel, G. A. Navratil, and C. Hansen. A dimensionality reduction algorithm for mapping tokamak operation regimes using variational autoencoder (VAE) neural network. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2020 of *APS Meeting Abstracts*, page NP16.007, Jan 2020.

[96] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[97] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[98] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[99] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[100] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[101] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[102] S. Dormido-Canto, J. Vega, J.M. Ramírez, A. Murari, R. Moreno, J.M. López, and A. Pereira. Development of an efficient real-time disruption predictor from scratch on JET and implications for ITER. *Nuclear Fusion*, 53(11):113001, 11 2013.

[103] J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, A. Pereira, and A. Acero. Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks. *Nuclear Fusion*, 54(12):123001, Dec 2014.

[104] DJ Campbell et al. The iter research plan. In *Proceedings of the 24th International Conference on Fusion Energy, San Diego, CA, USA*, pages 8–13, 2012.

[105] C. Rea, R. S. Granetz, K. Montes, R. A. Tinguely, N. Eidietis, J. M. Hanson, and B. Sammuli. Disruption prediction investigations using Machine Learning tools on DIII-D and Alcator C-Mod. *Plasma Physics and Controlled Fusion*, 60(8):084004, 8 2018.

[106] Cristina Rea and Robert S. Granetz. Exploratory Machine Learning Studies for Disruption Prediction Using Large Databases on DIII-D. *Fusion Science and Technology*, 74(1-2):89–100, 8 2018.

[107] C. Rea, K. J. Montes, A. Pau, R. S. Granetz, and O. Sauter. Progress toward interpretable machine learning-based disruption predictors across tokamaks. *Fusion Science and Technology*, 2020.

[108] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[109] George V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

[110] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[111] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[112] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[113] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.

[114] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3):121–136, 1975.

[115] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[116] Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. 04 1991.

[117] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[118] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[119] Kyunghyun Cho, Bart Merrienboer, Dzmitry Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. 09 2014.

[120] Shudong Yang, Xueying Yu, and Ying Zhou. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, pages 98–101, 2020.

[121] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[122] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. in eighth workshop on syntax. *Semantics and Structure in Statistical Translation (SSST-8), 2014*, 2014.

[123] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017.

[124] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[125] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[126] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[127] E. M. Hollmann, P. B. Aleynikov, T. Fülöp, D. A. Humphreys, V. A. Izzo, M. Lehnen, V. E. Lukash, G. Papp, G. Pautasso, F. Saint-Laurent, and J. A. Snipes. Status of research toward the ITER disruption mitigation system. *Physics of Plasmas*, 22(2):021802, 2 2015.

[128] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[129] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions*

on pattern analysis and machine intelligence, 12(10):993–1001, 1990.
[130] Simon Haykin. Neural networks: A comprehensive foundation, 3rd edn. 1999.
[131] MP Perrone and LN Cooper. [a]when networks disagree: Ensemble methods for hybrid neural networks, [o] neural networks for speech and image processing. *Chapman-Hall*, 1993.
[132] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.
[133] Jinxiang Zhu, Cristina Rea, RS Granetz, ES Marmar, KJ Montes, Ryan Sweeney, RA Tinguely, DL Chen, Biao Shen, BJ Xiao, et al. Scenario adaptive disruption prediction study for next generation burning-plasma tokamaks. *Nuclear Fusion*, 61(11):114005, 2021.
[134] K.J. Montes, C. Rea, R.A. Tinguely, R. Sweeney, J. Zhu, and R.S. Granetz. A semi-supervised machine learning detector for physics events in tokamak discharges. *Nuclear Fusion*, 61(2):026022, Jan 2021.
[135] A. Pau, A. Fanni, B. Cannas, S. Carcangiu, G. Pisano, G. Sias, P. Sparapani, M. Baruzzo, A. Murari, F. Rimini, M. Tsalas, and P. C. de Vries. A First Analysis of JET Plasma Profile-Based Indicators for Disruption Prediction and Avoidance. *IEEE Transactions on Plasma Science*, 46(7):2691–2698, 7 2018.
[136] Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks*, pages 195–201. Springer, 1995.
[137] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3):121–136, 1975.
[138] GL Jackson, TC Luce, WM Solomon, F Turco, RJ Buttery, AW Hyatt, JS DeGrassie, EJ Doyle, JR Ferron, RJ La Haye, et al. Long-pulse stability limits of the iter baseline scenario. *Nuclear Fusion*, 55(2):023004, 2015.
[139] RJ La Haye, C Chrystal, EJ Strait, JD Callen, CC Hegna, EC Howell, M Okabayashi, and RS Wilcox. Disruptive neoclassical tearing mode seeding in diii-d with implications for iter. *Nuclear Fusion*, 62(5):056017, 2022.
[140] E.J. Doyle, J.C. DeBoo, J.R. Ferron, G.L. Jackson, T.C. Luce, M. Murakami, T.H. Osborne, J.-M Park, P.A. Politzer, H. Reimerdes, Robert Budny, Thomas Casper, C.D. Challis, R.J. Groebner, C.T. Holcomb, A.W. Hyatt, R.J. Haye, G.R. McKee, T.W. Petrie, and Lunwu Zeng. Demonstration of iter operational scenarios on diii-d. *Nuclear Fusion*, 50:075005, 06 2010.
[141] Paul H. Rutherford. Nonlinear growth of the tearing mode. *Physics of Fluids*, 16:1903–1908, 1973.
[142] Roscoe B. White, D. A. Monticello, Marshall N. Rosenbluth, and B. V. Waddell. Saturation of the tearing mode. *The Physics of Fluids*, 20(5):800–805, 1977.
[143] R. J. La Haye, S. Günter, D. A. Humphreys, J. Lohr, T. C. Luce, M. E. Maraschek, C. C. Petty, R. Prater, J. T. Scoville, and E. J. Strait. Control of neoclassical tearing modes in diii–d. *Physics of Plasmas*, 9(5):2051–2060, 2002.
[144] R. J. La Haye. Neoclassical tearing modes and their control. *Physics of Plasmas*, 13(5):055501, 2006.
[145] Yousung Park, Steve Sabbagh, Jaeheon Ahn, B Park, Hyun-Seok Kim, John Berkery, Jim Bialek, Yanzheng Jiang, Jun-Gyo BAK, Alan Glasser, Jisung Kang, Jaehyun Lee, Hyunsun Han, Sang-Hee Hahn, Youngmu Jeon, B. Jung, Hyeon Kyun Park, Zhirui

Wang, Jong-Kyu Park, and S Yoon. Analysis of mhd stability and active mode control on kstar for high confinement, disruption-free plasma. *Nuclear Fusion*, 60, 02 2020.

[146] Karl Pearson F.R.S. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[147] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[148] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[149] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[150] R. Sweeney, A. J. Creely, J. Doody, T. Fülöp, D. T. Garnier, R. Granetz, M. Greenwald, L. Hesslow, J. Irby, V. A. Izzo, R.J. La Haye, N.C. Logan, K. Montes, C. Paz-Soldan, C. Rea, R.A. Tinguely, O. Vallhagen, and J. Zhu. MHD stability and disruptions in the SPARC tokamak. *Journal of Plasma Physics*, 86(5):865860507, 2020.