Automatic disruption classification in JET with the ITER-like wall

B Cannas¹, P C de Vries², A Fanni¹, A Murari³, A Pau¹, G Sias¹ and JET Contributors⁴

EUROfusion Consortium, JET, Culham Science Centre, Abingdon, OX14 3DB, UK
¹ ITER Organization, Route de Vinon sur Verdon, 13115 St. Paul lez Durance, France
² Consorzio RFX-Associazione EURATOM ENEA per la Fusione, I-35127 Padova, Italy

E-mail: alessandro.pau@diee.unica.it

Received 27 January 2015, revised 17 July 2015
Accepted for publication 14 August 2015
Published 16 October 2015

Abstract
The new full-metal ITER-like wall at JET was found to have a deep impact on the physics of disruptions at JET. In order to develop disruption classification, the 10D operational space of JET with the new ITER-like wall has been explored using the generative topographic mapping method. The 2D map has been exploited to develop an automatic disruption classification of several disruption classes manually identified. In particular, all the non-intentional disruptions have been considered, that occurred in JET from 2011 to 2013 with the new wall. A statistical analysis of the plasma parameters describing the operational spaces of JET with carbon wall and JET ITER-like wall has been performed and some physical considerations have been made on the difference between these two operational spaces and the disruption classes which can be identified. The performance of the JET- ITER-like wall classifier is tested in real-time in conjunction with a disruption predictor presently operating at JET with good results. Moreover, to validate and analyse the results, another reference classifier has been developed, based on the $k$-nearest neighbour technique. Finally, in order to verify the reliability of the performed classification, a conformal predictor based on non-conformity measures has been developed.

Keywords: disruptions, prediction, classification, machine learning, fusion plasma, real-time control

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

1. Introduction
The plasma disruptions in devices for the controlled thermonuclear fusion are associated to a sudden loss of magnetic confinement which causes the magnetic and thermal energy stored in the plasma to be released to surrounding structures. The consequent electromagnetic forces and heat loads have an impact on the operational lifetime of several components and in extreme cases seriously damage the integrity of the devices themselves. To date, the occurrence of disruptions has proven to be a hardly avoidable aspect of tokamak operation, particularly in high performance configurations; therefore, being able to understand them, on the route to the next step devices, it is a top priority to predict them, in order to undertake proper avoidance strategies and improve the efficiency of mitigation actions. For these reasons, the study of disruptive phenomena is particularly important. However, despite the intense study carried out during the last decades, a physical model that incorporates causes and occurrences of the disruptive instabilities is not yet available. Disruption phenomenology is extremely complex; disruptions have different causes and occur in different regions of the parameter operational space. These considerations motivate a strong
interest in implementing a disruption classifier, which, ana-
ylising the evolution of the plasma parameters, is able to dis-
minate among different disruptive behaviours by mapping
the discharge in different regions of the operational parameter
space. This allows not only to relate the plasma disruptivity to
the well-known operational boundaries, but also to investigate
over not straightforward relations in the high-dimensional
input space characterizing disruptive regions.

The work presented in this paper fits in the broad frame-
work of machine learning techniques that have been developed
as an original approach to automatic disruption classification
at JET. All the performed analyses are framed in the context of
disruption understanding. For example, a deeper knowledge
of the characteristics of the different disruption types can help
to plan countermeasures to avoid or mitigate an upcoming
disruption. Moreover, concerning the mitigation, it has been
proven in JET that the killer gas injection has not always the
same positive effect and it is imperative to understand whether
this depends on the disruption type, the disruption phase and
so on.

Machine learning methods have been extensively used in
the field of disruption prediction. In particular, several contri-
butions have been presented using neural networks (NNs) in
different tokamaks (Hernandez 1996, Wroblewski et al 1997,
Pautasso et al 2002, Sengupta 2001, Yoshino 2003, Cannas
et al 2004, 2007, 2010). Another successful experience in JET
is represented by the real-time advanced predictor of DISrump
tion (APODIS) (Vega et al 2013). In (Aledda 2012, Cannas
et al 2013a) the authors investigated the possibility of improving
the previous black box approaches, which are blind, or non-explan-
atory, by a process called manifold learning. Manifold learning
is based on the idea that there is more information in the data
than is exploited when using algorithms such as multi layer per-
ceptron neural networks or support vector machines. It finds
low dimensional structures in high dimensional data caused by
constraints on the data itself. Fewer efforts were made to apply
machine learning techniques to disruption classification.

The first attempt to automatically classify disruptions at JET
was described in (Cannas et al 2006) using pattern recognition
techniques. Disruptions for training were manually classified
in four classes by some of the authors, in collaboration with
physicists at JET. It has to be highlighted that, manually clas-
sifying disruption type is essential to develop any automated
classification system. In (Cannas et al 2013b) and (Murari
2013) both the proposed automatic disruption classifiers were
based on the manual classification proposed in (de Vries 2011)
for the discharges occurring during the JET operations with
the carbon wall (JET-C) from 2000 to 2010. In (de Vries 2011)
specific chains-of-events that led to disruption were identified
and used to classify disruptions, grouping those that follow
specific paths. Sometimes these paths are clear and unique,
while others could follow near similar courses. Moreover, sev-
eral different problems may occur simultaneously, eventually
leading to a disruption. Hence, not always an unambiguous
manual classification is possible. In (Cannas et al 2013b) the
potentiality of the generative topographic mapping (GTM) of
the JET-C operational space was exploited to develop an auto-
matic disruption classifier of seven disruption types classified
in (de Vries 2011), showing a great potential in terms of clas-
sification success rate (exceeding 97%).

The new full-metal ITER-like wall (JET-ILW) at JET
was found to have a deep impact on the physics of dis-
ruptions. Such impact has been analysed in (de Vries 2012, 2014) where it was stressed that the main difference
between JET-C and JET-ILW is the lengthening of the cur-
rent quench due to lower radiation and higher temperatures
during the disruption, which increases the impulse to the
vessel and conducts a larger fraction of energy to the wall.
This is aggravated by the fact that the ILW is more vulner-
able to heat loads.

Regarding the disruption causes, differences between
JET-ILW and JET-C were identified in (de Vries 2012, 2014)
for 2011 and 2012 campaigns. The predominant effects of the
ILW on disruption causes are the change in density limit, more
disruptions due to an error field locked mode, and a new class
of disruptions, due to accumulation of high-Z impurities. The
error field locked modes have become more common with the
JET-ILW because, in case of failure of the gas injection system,
the density could drop significantly allowing these modes to
grow, while with the JET-C the density would remain higher,
due to wall recycling. Accumulation of high-Z impurity was
observed in special cases with the JET-C. However, with the
JET-ILW it becomes the predominant disruption cause (de
Vries 2014).

In the present paper, a statistical analysis on JET-C and
JET-ILW disruptions has been performed to investigate how
the modification of disruption physics in the JET-ILW exper-
iments eventually influences the operational space of JET. The
analysis showed the necessity to develop a specialized GTM
map of the 10D JET-ILW plasma parameter space for disrup-
tion classification purposes. The results of the mapping have
been reported demonstrating the possibility to develop a real-
time application for the classification task in conjunction with
the APODIS disruption detection system currently working
online at JET (Vega et al 2013). Note that many methods and
analysis are often performed after the experimental campaigns

![Figure 1. 2D GTM of the 10D JET-C operational space (mode representation). The safe nodes are blue, the disruptive nodes are represented with different colours and symbols depending on the disruption class as indicated in the legend, empty nodes are white.](image-url)
(even several weeks), but the real challenge is the integration in the control systems of an automatic system that can be really able to recognize and detect disruptions with a sufficient warning time. The assessment of the real-time suitability of an automatic classification system is a consistent step toward this framework.

The potentiality of the proposed method in giving useful physics insight in the development of disruptions has been also discussed. In fact, even if the presented analysis is not explicitly aimed at providing any particular further physical understanding, the proposed classifier is able to give a deeper knowledge of the characteristics of different disruption classes supplying an image of the parameter space where each region is associated to one or more disruptive behaviours. Moreover, the developed automatic tools can be useful to determine the sequence of events leading to a disruption, which is a task very difficult to be manually performed and that could be very helpful to improve the disruption understanding and to optimize avoidance strategies.

Furthermore, in order to corroborate the obtained results, they have been compared with those obtained with a classifier based on $k$-nearest neighbour ($k$-NN). Finally, in order to verify the reliability of the classification, a conformal predictor has been developed which provides information on the level of confidence of the proposed classification.

2. Generative topographic mapping of the JET-C disruption operational space

Generative topographic mapping belongs to the class of the so-called ‘generative models’, which try in various ways to model the distribution of the data by defining a density model with low intrinsic dimensionality in the data space. Through a nonlinear mapping from the latent space to the data space, the GTM generates a mixture of Gaussians, whose centres are constrained to lie on a low-dimensional space embedded in the high-dimensional one, which has to be fitted to the data. This is usually achieved through a form of the expectation maximization algorithm by maximizing the likelihood or the log-likelihood function of the model (Bishop 1998). GTM can be applied for data clustering and topology preservation. Being the mapping defined by a smooth and continuous nonlinear function, the topographic ordering of the latent space

Figure 2. Kernel probability density estimates of: (a) plasma current ($I_p$); (b) safety factor at 95% of poloidal flux ($q_{95}$); (c) plasma internal inductance ($l_i$); (d) line integrated plasma density ($n_e$).
will be preserved in the data space, in the sense that points close in the latent space will be mapped onto nodes still close in the data space.

In (Cannas et al. 2013b) the GTM of the 10D operational space of JET-C was used to develop a disruption classifier of seven disruption classes manually classified in (de Vries 2011). In particular, 243 non-intentional disruptions occurred on JET in the experimental campaigns from 2005 up to 2009, in the shot range between 63,718 and 79,853, were considered. In the aforementioned interval, 1467 safe discharges were also selected. The plasma quantities were: the plasma current ($I_p$); the poloidal beta ($\beta_p$); the mode lock amplitude signal ($LM$); the safety factor at 95% of poloidal flux ($q_{95}$); the total input power ($P_{tot}$); the plasma internal inductance ($l_i$); the plasma centroid vertical position ($Z_{cen}$); the line integrated plasma density ($n_{eold}$); the stored diamagnetic energy time derivative ($dW_{diamag}/dt$); the total radiated power ($P_{rad}$). The majority of the chosen quantities refers to signal amplitudes rather than signal derivatives because the focus has been put on the description of the parameter operational space of JET disruptions (Cannas et al. 2013a). Each signal was sampled at 1 kHz, and a ‘safe’ label was associated with each sample of the safe discharges whereas a ‘disrupted’ label was associated with the last 210 ms of the disruption terminated discharges (one sample every 1 ms in the time interval $[t_D-210-t_D]$ms, where $t_D$ is the disruption time (Cannas et al. 2013a)). Then, a data reduction was performed for the safe discharges to reduce the huge amount of safe samples and to balance the data set of safe and disrupted samples.

In de Vries (2011) the non-intentional disruptions in the considered JET-C campaigns were analysed and associated with particular disruption classes by detecting specific chains of events and grouping those that follow definite paths. In particular, the following seven classes were identified: problems during the auxiliary power shut-down (ASD), greenwald limit (GWL), impurity control problem (IMC), internal transport barrier (ITB), low density and low q (LON), density control problem (NC), neo-classical tearing mode (NTM). It should be noted that the complexity of the disruption process could make this manual classification rather ambiguous and a few disruptions were not able to be classified at all (de Vries 2011). Nevertheless, this work was essential to develop an automated classification able to help identifying a strategy for disruption avoidance.

Making reference to the manual classification, each disruptive sample in the last 210 ms of the discharge is labelled with the disruption type. Note that, the sequence of events used to define the disruption type does not necessarily occur in the time interval $[t_D-210, t_D]$ms of the considered disruption. For example, a disruption could be classified into NTM type according to the initial phenomenology which deteriorates the plasma, while it disrupts for MARFEs which characterizes another class. However, what we want to evaluate here is how the different classes influence the final evolution of the discharge and the corresponding parameter space.

### 2.1 Automatic classification of the JET-C disruptions

In figure 1, the 2D GTM of the 10D JET-C operational space is reported, making reference to the mode representation (Bishop 1998). In the GTM, the latent space is a discrete grid of nodes (or cells). The arrangement of nodes is a 2D regular spacing in a $70 \times 70$ rectangular grid. Each map unit in the GTM can be associated with a particular composition characterized by a coloured symbol, as shown in the legend in figure 1.

Such mapping technique has been also exploited for disruption classification purposes trying to retrieve clusters in the map corresponding to different disruption classes. Figure 1 shows that some classes are quite widespread all over the disruptive regions, but also regions where a specific class results to be predominant with respect to the others can be found. Thus, there is not only a well-defined separation between disruptive and non-disruptive regions, but also the possibility to characterize certain regions with a higher probability for a certain class with respect to the others. For example, it can be seen that disruptions due to too strong ITBs, which are characterized by a well-defined physics, are projected in the lower right corner of the GTM, while several regions mostly contain both NCs and IMCs.

As previously mentioned, each node in the map is related to samples coming from different classes. By projecting onto the map the temporal evolution of a discharge, each sample results to be associated with a node. For each sample and each class, a class membership can be defined on the base of the percentage of samples of the considered class in the

### Table 1. GTM success rates of the automatic disruption classification for JET-C.

| Class       | GLOBAL | ASD | GWL | IMC | LON | NC | NTM | Success rate [%] |
|-------------|--------|-----|-----|-----|-----|----|-----|------------------|
|             | 97     | 100 | 100 | 99  | 100 | 100| 92  |                  |

### Table 2. Composition of the JET-C and JET-ILW non intentional disruption databases according to the manual classification.

| Disruptions         | JET-C | JET-ILW |
|---------------------|-------|---------|
| Labels              | Classes | Num | Num % | Num | Num % |
| ASD                 | Auxiliary power shut-down | 50 | 20,58 | 2 | 1,34 |
| GWL                 | Greenwald limit            | 9 | 3,70 | 0 | 0,00 |
| IMC                 | Impurity control problem   | 83 | 34,16 | 28 | 18,79 |
| IMC_high-Z          | New impurity control problem | 0 | 0,00 | 82 | 55,03 |
| ITB                 | Internal transport barrier | 10 | 4,12 | 0 | 0,00 |
| LON                 | Low density and low q      | 12 | 4,94 | 7 | 4,70 |
| NC                  | Density control problem    | 58 | 23,87 | 22 | 14,77 |
| NTM                 | Neo-classical tearing mode | 21 | 8,64 | 8 | 5,37 |
node to which the sample is associated, with respect to the
total number of disruptive samples in the node itself. In order
to classify a disruptive shot, a majority voting algorithm has
been adopted based on the class membership of each class in
a prefixed time interval before the disruption. The automatic
classification has been performed in the last 210 ms of the
disrupted pulses and it was in very good agreement with the
manual classification, as reported in table 1.

2.2. JET-ILW versus JET-C disruption operational spaces

After the installation of the new ILW it was first attempted to
project the disruptions of the JET-ILW campaigns onto the
GTM trained with the JET-C discharges. The performance
of the map in classifying the new disruptions significantly
deteriorated for certain classes (especially for IMC), prob-
bly because of the fact that the operational space or, at least,
the considered feature space changed. Therefore, a detailed
analysis has been performed to investigate how the modifi-
cation of the disruption physics in the JET-ILW experiments
with respect to the JET-C ones, recognised in de Vries (2012),
eventually influences the classification space of JET.

As mentioned in the introduction, the most common
disruptions during the first phase of operation with the
ILW, were those due to accumulation of high-Z impurities
(mainly W), and as a consequence, excessive core radiation.
Originally, for JET-C operations, the class IMC was pro-
posed to deal with disruptions due to impurity control prob-
lems. However, for JET-ILW operations it was found that a
new distinct class existed, related particularly to the control
of high-Z impurities (de Vries 2014), which in this paper is
labelled IMC_high-Z.

In the data set for the JET-C wall classifier the IMC disrup-
tions were mainly due to low-Z impurities and linked to large
electrode radiation. This resulted in the shrinking of the plasma
column, yielding the growth of instabilities that disrupted the
plasma. Conversely, the new IMC_high-Z class has features
that are quite distinct from the IMC class, such as accumula-
tion of high-Z material, strong core radiation and the formation
of hollow temperature profiles, which result in the flattening of
the current density profile, yielding again an onset of instabili-
ties (de Vries 2014). Such disruptions were rare with the JET-C
and hence previously not identified as a separate class (de Vries
2011). The root cause of the disruptions due to high-Z impu-
rity accumulation may lie in the edge, where sputtered material
enters the plasma, although a clear cause is not often found.

For these reasons, disruptions related to high-Z impurity con-
trol have been considered separately as a new class.

To evaluate whether this modification in the physics of the
disruptions has changed the disruption operational space, a sta-
tistical analysis has been performed on JET-C and JET-ILW
disruption classes. For JET-C the database consists of 243 non
intentional disruptions occurred between 2005 and 2009; for
JET-ILW it consists of 149 non intentional disruptions that
occurred between 2011 and 2013. In table 2 the distribution and
the occurrence for the different classes for both the JET-C and
the JET-ILW are reported. As it can be seen, the composition of
the two databases is quite different: in particular, disruptions due
to the Greenwald limit or due to too strong ITB are no longer
present in the new campaigns, whereas the new IMC_high-Z

Figure 3. Kernel probability density estimates of $I_p$ (left side) and $I_i$ (right side) for the IMC (grey) and NC (green) disruptions with JET-C.

Figure 4. Kernel probability density estimates of $I_p$ (left side) and $I_i$ (right side) for the IMC (dashed grey), IMC_high-Z (dashed blue) and NC (dashed green) disruptions with JET-ILW.
becomes the predominant class, as earlier reported (de Vries 2014). Note that, the relative percentage of each class of disruption is strongly influenced by the scientific program, and/or by the number of sessions devoted to a given program.

In figure 2 the kernel probability density estimates (kpde) of four plasma parameters related to the last 210 ms of the IMC disruptions for the JET-C (red lines) and JET-ILW (grey dashed lines), and IMC_high-Z for JET-ILW (blue dashed lines) are reported. The parameters are: (a) plasma current $I_p$; (b) safety factor at 95% of poloidal flux $q_{95}$; (c) plasma internal inductance $l_i$; (d) line integrated plasma density $n_{ei}$. The analysis gives us interesting information. In particular, the kpdes in figure 2 show operational spaces and parameter ranges of JET-C quite different from JET-ILW. Moreover, very few disruptions due to high-Z control problems occurred at JET-C, whereas they are the predominant cause of disruptions at JET-ILW. This analysis confirms that a new GTM is needed to represent the JET-ILW operational space. Furthermore, from figure 2 it can be seen that it is quite difficult to discriminate among classes by the distribution of the signals. In fact, it is well known that what really matters is the combination of the parameters.

For the IMC_high-Z class the kpde of the internal inductance is shifted towards lower values, whereas the kpde of the electron density is shifted towards higher values. This is a direct indication of the impact of the high-Z material on the core density and its radiation, flattening the current density profile, thus lowering the internal inductance.

Further analyses can be performed to compare different behaviour of disruption classes passing from JET-C to JET-ILW. Regarding the density control problem and the impurity control problem classes, figure 3 reports the kpdes of $I_p$ and $l_i$ for the last 210 ms of IMC and NC disruptions with JET-C, whereas figure 4 reports the distributions of the same signals for the IMC, IMC_high-Z and NC disruptions with JET-ILW. From figure 4, it can be seen that, for the JET-ILW, both $I_p$ and $l_i$ signals result to be quite different, especially if NC and IMC_high-Z classes are compared. In particular, the new impurity control problem type basically occurs for lower values of plasma inductance, mainly as a result of the flattening or the hollowing of the current profiles. Regarding the plasma current, it can be seen that no NC disruptions occur above 2 MA: note that the high values of $I_p$, in the case of IMC_high-Z disruptions, are probably due to the typical ranges of currents used in the attempt to control high-Z impurity accumulation. Therefore, in this case, the distributions are showing the statistical evidence of the considered databases and not a direct dependence of high values of $I_p$ with high-Z impurities. Conversely, the JET-C, NC and IMC disruptions share the same region in the operational space (Cannas et al 2013b). This is confirmed also looking at figure 3, where the kpdes of $I_p$ and $l_i$ are more or less overlapped.

3. Mapping of the JET-ILW disruption operational space

Starting from the previous statistical analysis and the physical considerations on the new disruption class, a new GTM has been trained to represent the JET-ILW operational space. The training set consists of the last 210 ms of the 149 non intentional JET-ILW disruptions (29 137 samples). For the sake of comparison, the same plasma parameters used to represent the JET-C has been considered; the resulting GTM has a latent space of $36 \times 36$ grid of nodes built using 81 radial basis functions (Gaussian shape) with a 1.5 width. In figure 5(a) the mode representation of the GTM is reported. Figure 5(b) shows the GTM pie plane representation. In such visualization, each node is represented by a pie chart describing the percentage composition in terms of number of samples belonging to the different classes. The samples are diversified according to the colour code reported on the legend. Both representations highlight a high level of separation among the different classes.

The disruption types that have been taken into account are characterized in some cases by a well-defined physics, in
other cases are mainly related to technical and control problems. Nevertheless, as discussed in the previous work (Cannas et al. 2013b), there are cases, such as NC and IMC, in which clearly distinguishable problems give rise to a similar phenomenology in the final phase (radiative collapse and cooling of the edge). However, from the avoidance and the machine operation point of view, it could be useful to know when the discharge is evolving in a region of the considered parameter space with high risk of disruption in relation to specific operations (switch off of the auxiliary power, impurity seeding, etc.). In this framework, a characterization of the operational space able to discriminate among such problems would also be helpful to think about different control strategies aiming at minimizing the disruptivity related to certain scenarios/operations. In particular, from figure 5, it can be seen that the new IMC_high-Z class is well separated with respect to the other classes. The percentage of samples of the IMC_high-Z class, which is projected in nodes entirely composed by samples of the same class, is more than 91% and the remaining samples overlap mainly with the IMC class. Further useful information can be obtained by looking at the component planes of some more significant signals.

The component plane representation expresses the relative component distribution of the input data on the 2D map (Cannas et al. 2013a), allowing to visually identify eventual similar patterns or particular behaviours for certain classes. As an example, the differences in terms of the plasma current and the internal inductance for the density control problem (green dots) and the high-Z impurity control problem (blue dots) classes can be pointed out by analysing the corresponding component planes shown in figure 6.

Similar considerations to those made for the kpdes of figure 4 can be done: in particular, it is easy to see how density control problem disruptions occur for low values of the \( I_p \) and high values of the \( l_i \). Conversely, most of the impurity control problem disruptions occur for higher values of the \( I_p \) and lower values of the \( l_i \). Note that, IMC_high-Z disruptions occur for a broader \( I_p \) and \( l_i \) distribution but with different frequencies. These tools, together with the statistical analysis, can provide efficiently non-trivial information of a complex multidimensional space, which usually is quite hard to attain with classical methods.

4. Automatic classification of the JET-ILW disruptions

In order to test the classification performance of the JET-ILW map, the majority voting algorithm has been applied to the last
210 ms of each discharge: all but one of the disruptions have been correctly classified according to the manual classification, confirming the GTM discrimination capability among the different classes.

One of the main objectives of this study is to develop a system that can be used in real time and can be eventually integrated with the other systems already working in real time at JET. Therefore, a real time application has been simulated in conjunction to APODIS: the majority voting algorithm has been applied to the class membership function in a time window of 32 ms right before the time in which APODIS triggers the alarm. The choice of such a time window is in accordance with APODIS itself, which generates the alarm through a decision function based on the behaviour evaluated using time windows of 32 ms. Note that, in several cases APODIS gives the alarm significantly in advance with respect to the thermal quench, even hundreds of ms in advance.

Table 3 reports the performance in terms of success rate of the real time automatic classification achieved by the GTM. The results refer to the 149 considered disrupted pulses. A pulse has been considered correctly classified if the automatic system produces the same classification given in de Vries (2014). Hence, the success rate is the percentage of disruptions correctly classified. As it can be seen, the global success rate is quite high, reaching 87%, hence in good agreement with the manual classification. Some problems arise from the difficulty to discriminate the IMC class from the IMC_high-Z control problem one, at least on the base of the selected plasma parameters. Other signals, such as core radiation or radiation peaking, should be included to better discriminate the two IMC classes, but such signals are not always reliable for all the disruptions in the database.

4.1. Disruption classification by k-NN

In order to validate and analyse the results obtained with GTM, another reference classifier has been developed based on k nearest neighbour (k-NN), which uses as kernel the Mahalanobis distance (Mahalanobis 1936). It represents one of the simpler but at the same time more used classification algorithms. An object can be classified on the base of its neighbours classification; the object is assigned to the class with the higher number of neighbours among the k nearest ones. k-NN is defined as an instance-based classifier, unlike GTM for example, which defines a generative latent model. k-NN is a simple and flexible technique whose drawbacks are well known, as for example the application of the majority-voting criterion for classification when the dataset is strongly unbalanced in terms of the different classes. However, the method has some strong consistency results. In particular, the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error if the amount of data tends to infinity (Cover 1967). The Bayes error gives a statistical lower bound on the error achievable for a given classification problem and associated data Duda et al (2000).

In the present case, the majority voting is applied to the k closest points in the tokamak high dimensional space and it can be interpreted also in terms of Bayes’ formalism. Table 4 reports the performance of the k-NN classifier. The global performance is above 90% when a time window of 32 ms before the APODIS alarm is considered. These results are even better than those of GTM. However, in the problem at the hand, the huge amount of points to be considered at the same time prevents the application of the method in real time. In order to shorten the k-NN execution time, the search process could be parallelized and/or optimized and high-dimensional similarity.
search techniques might be used (Broder 1986, Nene 1997). As different classifiers may offer complementary information about the patterns to be classified, thus combining classifiers could provide better classification results than any single classifier, even the best one. Presently, $k$-NN can be used off-line, in parallel with the GTM classifier combining $k$-NN and GTM outputs in an efficient way.

4.2. Class-membership function

The temporal sequence of the samples of a disruption (the last 210 samples of the disrupted shots), can be projected on the GTM map associating each sample with one node. For each sample and each class, the class membership is evaluated, being the percentage of samples of the considered class in the node to which the sample is associated, with respect to the total number of disruptive samples in the node itself. The temporal evolution of the class membership of each class is called class-membership functions. The class-membership function gives us useful information. As an example, in figure 7 the class-memberships of the JET-ILW disruption No. 82867 is reported for both GTM and $k$-NN; it results to be an IMC disruption according to the manual classification. It is possible to note a transition among different classes and in particular that between NCs and IMCs or vice versa, which is not uncommon both for JET-C and JET-ILW disruptions. It means that the characteristics of the disruption process change in time, and are detected differently long before the disruption and closer to the disruption time. Note that APODIS alarm is triggered almost two seconds before $t_D$, whereas in the $k$-NN more than 700 ms before $t_D$.

In figure 8 the time evolution of some of the available signals is reported for the same discharge (No. 82867) with reference to the time window analysed in figure 7. As can be seen in figure 8, a locked mode grows at $t = 13.79$ s, around which a rapid change of the density occurs, followed by a quench of the temperature that, in the subsequent phases, recovers up to the final thermal quench at $t = 15.73$ s. Both PTN and APODIS trigger the alarm when the mode locks (see figure 7) and for both classifiers the discharge evolves as a NC disruption up to the final phase where is correctly classified as IMC, according to the manual classification.

4.3. Conformal predictors

Given the complex behaviours that often characterize the evolution of a discharge, it is important to know the reliability of the classification. The literature provides recent methods, such as the conformal predictors, which allow us to take into account also this aspect (Shafer 2008, Vovk 2010). To this purpose, a conformal predictor has been developed which is based on non-conformity measures. Note that, unlike others’ methods, conformal predictors have the peculiarity to provide, together with prediction or classification, the corresponding level of confidence, even if the well-known constraints related to the computation time restrict their application in real time.

The theory of conformal predictions is based on the principles of algorithmic randomness, and on the Kolmogorov complexity of an IID (identically independently distributed) sequence of data instances. Conformal predictors can be used together with any prediction method, such as support vector machines, neural networks, decision trees, or nearest
neighbour classifiers. Recently, a method based on membership functions has been proposed to extend their use also to fuzzy logic classifiers (Murari 2015). To determine the confidence level for the classification of a new object, it is necessary to estimate how different the new object is from the old examples: to this purpose, a nonconformity score is usually calculated on the base of a defined nonconformity measure.

Let us consider $n$ successive ordered pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $z_i = (x_i, y_i)$ represents the generic example, which consists of an object $x_i$ and the corresponding label $y_i$. Both the object and the label belong to measurable spaces, respectively the object and the label space. A bag of size $n$ is a collection of $n$ elements and can be given in any order. In the following, a bag of size $n$ will be indicated with the notation $(z_1, \ldots, z_n)$. The first step of the conformal prediction algorithm is the computation of the nonconformity scores $\alpha_i$ for any object of the given bag on the base of a defined nonconformity measure $A$:

$$\alpha_i := A(z_1, \ldots, z_i-1, z_{i+1}, \ldots, z_n, z_i)$$

Nonconformity scores have not an absolute value, being relative to the particular case considered for the given bag of objects $(z_1, \ldots, z_n)$. Therefore, in order to generalize and give a measure of how unusual an element $z_i$ is with respect to the other elements of the bag, its score must be compared with the one of all the other objects. This can be done for example by computing the so-called $p$-value, which is defined by the fraction of objects for which the non-conformity score is greater than $\alpha_i$:

$$p\text{-value}_i = \frac{\# \{\alpha_j \geq \alpha_i \}_{j=1,\ldots,n}}{n}$$

This fraction, which is the $p$-value for $z_i$, can assume values between 1/$n$ and 1, and represents the normalized number of examples belonging to the bag at least as nonconforming as $z_i$. The closer to its lower bound 1/$n$ the $p$-value is the more nonconforming the object $z_i$ is with respect to the other elements of the bag. If $n$ is large enough, a high level of nonconformity may define an outlier for the considered class.

In the framework of the classification with conformal predictors, the $p$-values have a double function: they are used to assign the class to a new element and, at the same time, on the base of their values, it is possible to define the goodness and the reliability of the classification itself. Thus, for a new object of an unknown label, which has to be classified on the base of the defined nonconformity measure into one of the available classes, the conformal predictor will assign to it the label with the highest $p$-value. The reliability of the prediction is quantified by two parameters: credibility, which is the largest $p$-value in the bag ($\text{Credibility} = \max_j (p\text{-value}_j), j = 1, \ldots, n$), and confidence, which is the $1$’s complement of the second largest $p$-value (Confidence = $1$–2nd largest $p$-value).

Assuming that each class is statistically well represented in the training set, the values of credibility and confidence are indicative of the reliability with which the classification is provided. In particular, the credibility is a measure of how the object (test) is representative of the assigned class: a low value of credibility means that the test is not representative of any class in the training set. The confidence is related to the ambiguity of the classification: it measures how close to each other the first two classifications are. Another important point is represented by the fact that the maximum $p$-value is not necessarily defined in a unique way, in the sense that the maximum $p$-value could be attributed to more than one class. This is a case of ambiguity, which means the conformal predictor for the given training set, on the base of the defined nonconformity measure, is not able to discriminate among the classes with which the maximum $p$-value is associated.

Conformal predictors measure the nonconformity of the new example $z_n = (x_n, y_n)$ with respect to the old ones belonging to the training set quantifying the goodness of the prediction. The nonconformity score can be computed in different ways. For the classification purpose of this work, the conformal predictor implements, for the measure of nonconformity, the nearest neighbour technique. In particular,
assigning $x_i$ to all the possible classes, for each object $x_i$ the implemented conformal predictor compares its distance from the nearest object with the same label with its distance from the nearest neighbour with a different label computing the so-called nonconformity score:

$$\alpha_i = \frac{\min \{ \|x_i - x_j\| : y_i = y_j \}}{\min \{ \|x_i - x_j\| : y_i \neq y_j \}}$$

In figure 9 the class membership provided by the classifier is reported together with the credibility and the confidence levels for the JET-ILW disruption No. 82 867. As can be seen, the credibility, which is the parameter with more variability, is quite low for all the initial phase, and then it rises constantly during the last 400 ms, according to the results obtained with the GTM based classifier. The credibility, even if low in the phase where the conformal predictor assigns the label corresponding to the NC class, is mostly above 0.05, which in the literature (Gammerman 2007) is often used as threshold for trusting or not a prediction (right side of figure 9). In general, if the credibility is less than 5%, the considered samples are not representative of the training set, or in other words,
they cannot be considered as generated independently from the same distribution. In particular, the credibility falls under the considered threshold in the correspondence of the transition between NC and IMC classes. This behaviour could depend on a rapid reconfiguration or a change in the considered parameters’ space. Further analysis should be done to clarify this point.

In figure 10, the class membership function obtained with the GTM (a) and with the \(k\)-NN (b) based classifiers are reported for the JET-ILW disruption No. 82569, which has been manually classified as IMC disruption.

It can be noted that, in addition to the agreement in the classification provided by the two methods, the confidence level plotted in figure 11 remains very high and almost stable up to the last 100 ms before the disruption. This behaviour is indicative of the fact that, in this phase, the considered plasma parameters are not changing, some of which have been reported see figure 12. That is coherent with the projection on the GTM map (see figure 13), where the discharge is evolving in a limited region of the operational space.

In figure 14 an example of a JET-ILW discharge (No. 82669) disrupted due to impurity accumulation is shown, i.e. the IMC_high-Z class. In this case, the accumulation of \(W\) occurs after a step-down of the neutral beam injection power (de Vries 2012), and the hollowing of the temperature profile can be observed. Eventually the instabilities that are triggered by the broadening of the current density profile lock and a disruption takes place. Figures 15 and 16 report the class-membership functions calculated through the GTM and the \(k\)-NN classifiers, and through the conformal predictor, respectively, for the aforementioned pulse. The three predictors classify the pulse as IMC_high-Z according to the manual classification. Furthermore, it is interesting to see that when the mode locks there are ‘jumps’ in the class membership calculated by the conformal predictor, and the credibility in the corresponding time interval drops almost to zero. In the interval prior to the locked mode, again the three classifiers clearly recognize the new impurity type. Note that, for all the misclassified disruptions the three classifiers are in agreement, confirming the robustness of the proposed methods. The misclassifications
Figure 14. Example of disruption caused by impurity accumulation (JET-ILW discharge No. 82 669).

Figure 15. Class-membership functions calculated through (a) GTM, (b) \(k\)-NN for the JET-ILW discharge No. 82 669.

Figure 16. (a) Class-membership functions calculated for the JET-ILW disruption No. 82 669 through the conformal predictor; in (b) a zoom of (a) is reported regarding the confidence level (black) and the credibility (blue).
5. Conclusions and discussion

The problem of automatically discriminating the type of disruption at JET both in the carbon wall and in the ITER-like wall campaigns has been investigated using a GTM manifold learning method. The disruption classes in the JET-ILW have been deeply analysed and compared with those in the JET-C. In particular, the kernel probability density estimates of the different plasma parameters highlight the different behaviours of the new impurity control disruptions, due to tungsten accumulation in the core of the plasma column, with respect to the old impurity control disruptions. Moreover, the statistical analysis showed the variation of the JET-ILW operational space with respect to that with JET-C. Note that, the disruption classes may be specifically defined also in relation to the specific operation and characteristics of the machine, and they are not supposed to be universally defined for every existing machine, even if, of course, some of them can be found in other machines with similar characteristics. For this reason, a new GTM map has been trained for JET-ILW. The latter has been used to simulate a real time behaviour of the GTM classifier in conjunction with the prediction system APODIS, which is successfully working on line at JET. The obtained results assess the suitability of the GTM based classifier for real time applications with good results: the prediction success rate is quite high (87%) according to the manual classification. The good results motivate the deployment of this tool in the real time digital network (ATM) of JET. Furthermore, in order to validate and analyse the obtained results, another reference classifier has been developed based on k-NN that uses as kernel the Mahalanobis distance. The performance of the reference classifier is above 90%, but, the huge mass of the data to be considered prevents the use of the method in real time applications. In particular, the calculation time required by the two methods has been assessed on a portable workstation of average performance. GTM based code takes an average time/sample of ~0.7 ms, giving us the possibility to stay within a reasonable time resolution of 1 ms, whereas k-NN based code takes an average time/sample of ~2.76 ms. Moreover, the calculation time of the latter would increase linearly with the size of training data, making difficult its exploitation in real-time with the aforementioned time resolution.

Several visualization tools have been developed for the GTM, such as pie plane representation or component plane representation, which make possible to extract relevant information that confirms the physical characteristics of the different classes. Monitoring the evolution of each disruptive discharge on the GTM, a class membership has been defined by means of which it is possible to perform a statistical analysis of the transitions among different classes.

Finally, in order to verify the reliability of the performed classification, a conformal predictor has been developed which is based on non-conformity measures. The obtained results indicate the suitability of the conformal predictors to assess the reliability of the GTM classification even if the computational time allows their use only in an off line fashion.

Summarizing, the exploitation of the developed tools to provide physics insight of a complex multidimensional space is very promising for several reasons. The GTM model, unlike k-NN and conformal predictors, can be exploited for data visualization purposes (Cannas et al 2013a, 2013b), allowing advanced analysis of the operational space where the relevant physics takes place. The tool allows us to uncover structures hidden in the high dimensional data studying and identifying dependencies and relations among different features by analyzing similar patterns in data. In fact, the projection on the 2D map is connected through the component planes to the relative distribution of each input parameter. This provides the possibility to look simultaneously at different planes. From one hand, the evolution of an operative point on the 2D map provides the information related to the performed classification (safe/disruptive, different disruption classes, etc). On the other hand, the operative point evolution with respect to the component distribution of the input parameters allows consideration on the position with respect to certain operational boundaries. For this purpose, a future activity will regard the development of further functionalities of the tool, such as the calculation of the distance between different clusters or regions on the map in relation to the distances in the data space. These functionalities, besides providing an indication about the embedding of the manifold in the high dimensional space, could be useful to have an indication about how far an operational point is from a given boundary. In addition, a better characterization of sudden changes in the plasma parameters space or transitions among different states during the evolution of a discharge will be possible. Therefore, the tool developed in this work may have several applications also in the framework of plasma control and operations.

The analysis described in this paper has several additional values, but it is worth discussing some potential criticism concerning specifically the classification and, in a more general contest, the prediction of disruptions. Even if many of the physics processes and criteria that drive the classification can be generalized to a certain extent, one has to keep in mind that differences in geometry, configuration, material, etc, can affect common physics processes leading to disruption. Therefore, to build a ‘universal’ classifier is unfeasible; nevertheless, in the view of ITER, the previous experience on machines with similar characteristics can be exploited to extrapolate some knowledge to the new machine. Furthermore, some disruption classes are characterized by common processes or intrinsic limits, whose dependence on the specific machine characteristics can be dealt with taking into account dimensionless parameters.

Any data-based disruption classifier or predictor needs to a certain extent a training phase and this represents a serious concern for the applicability to ITER, where unmitigated disruptions, in general, will not be allowed. However, it is worth mentioning that ITER, in the initial phase, will not operate at

Finally, in order to verify the reliability of the performed classification, a conformal predictor has been developed which is based on non-conformity measures. The obtained results indicate the suitability of the conformal predictors to assess the reliability of the GTM classification even if the computational time allows their use only in an off line fashion.

Summarizing, the exploitation of the developed tools to provide physics insight of a complex multidimensional space is very promising for several reasons. The GTM model, unlike k-NN and conformal predictors, can be exploited for data visualization purposes (Cannas et al 2013a, 2013b), allowing advanced analysis of the operational space where the relevant physics takes place. The tool allows us to uncover structures hidden in the high dimensional data studying and identifying dependencies and relations among different features by analyzing similar patterns in data. In fact, the projection on the 2D map is connected through the component planes to the relative distribution of each input parameter. This provides the possibility to look simultaneously at different planes. From one hand, the evolution of an operative point on the 2D map provides the information related to the performed classification (safe/disruptive, different disruption classes, etc). On the other hand, the operative point evolution with respect to the component distribution of the input parameters allows consideration on the position with respect to certain operational boundaries. For this purpose, a future activity will regard the development of further functionalities of the tool, such as the calculation of the distance between different clusters or regions on the map in relation to the distances in the data space. These functionalities, besides providing an indication about the embedding of the manifold in the high dimensional space, could be useful to have an indication about how far an operational point is from a given boundary. In addition, a better characterization of sudden changes in the plasma parameters space or transitions among different states during the evolution of a discharge will be possible. Therefore, the tool developed in this work may have several applications also in the framework of plasma control and operations.

The analysis described in this paper has several additional values, but it is worth discussing some potential criticism concerning specifically the classification and, in a more general contest, the prediction of disruptions. Even if many of the physics processes and criteria that drive the classification can be generalized to a certain extent, one has to keep in mind that differences in geometry, configuration, material, etc, can affect common physics processes leading to disruption. Therefore, to build a ‘universal’ classifier is unfeasible; nevertheless, in the view of ITER, the previous experience on machines with similar characteristics can be exploited to extrapolate some knowledge to the new machine. Furthermore, some disruption classes are characterized by common processes or intrinsic limits, whose dependence on the specific machine characteristics can be dealt with taking into account dimensionless parameters.

Any data-based disruption classifier or predictor needs to a certain extent a training phase and this represents a serious concern for the applicability to ITER, where unmitigated disruptions, in general, will not be allowed. However, it is worth mentioning that ITER, in the initial phase, will not operate at
full performance, therefore the disruption classifier and prediction systems can somehow be trained. Of course, this will not answer all the questions exhaustively, but it could provide a first insight in terms of the dominant processes leading to disruption in relation to the explored regimes. Then, if the transition to full performance regimes are gradually carried out, a step-by-step adaptive retraining could be performed without starting from scratch.

Concerning technical aspects, future work will be devoted to integrate and refine the proposed approach by considering different weights for certain parameters on the base of conditions or rules to be defined through both physical and statistical considerations. Such integration, with the introduction of constraints, could be fundamental to taking into account any additional information such as stability limits. This would give rise to the ‘supervision’ of an unsupervised system through physics and statistics.

Reliable algorithms for disruption avoidance and prediction are foreseen as part of the ITER plasma control system. The obtained results demonstrate that, even if the developed classifiers are machine dependent, the methodology is general. The next study will be oriented in the direction of finding a set of dimensionless parameters in order to assess the portability of these machine learning methods and analysis on different machines, but this first step was needed to assess the potentiality of the methods themselves.

Acknowledgment

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014–2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

References

Aledda R, Cannas B, Fanni A, Sias G, Pautasso G and the ASDEX Upgrade Team 2012 Mapping of the ASDEX upgrade operational space for disruption prediction IEEE Trans. Plasma Sci. 40 570–6
Bishop C, Svensén M and Williams C 1998 GTM: the generative topographic mapping Neural Comput. 10 215–34
Broder A J 1986 Strategies for efficient incremental nearest neighbor search Pattern Recognit. 23 171–8
Cannas B et al 2004 Disruptions forecasting at JET using neural networks Nucl. Fusion 44 68–76
Cannas B et al 2007 A prediction tool for real-time application in the disruption protection system at JET Nucl. Fusion 47 1559–69
Cannas B, Cau F, Fanni A, Sonato P, Zedda M K and JET-EFDA contributors 2006 Automatic disruption classification at JET: comparison of different pattern recognition techniques Nucl. Fusion 46 699–708
Cannas B, Fanni A, Murari A, Pau A, Sias G and JET EFDA Contributors 2013a Manifold learning to interpret JET high-dimensional operational space Plasma Phys. Control. Fusion 55 045006
Cannas B, Fanni A, Murari A, Pau A, Sias G and JET EFDA Contributors 2013b Automatic disruption classification based on manifold learning for real time applications on JET Nucl. Fusion 53 095023
Cannas B, Fanni A, Pautasso G, Sias G and Sonato P 2010 An adaptive real-time disruption predictor for ASDEX Upgrade Nucl. Fusion 50 075004
Cover T M and Hart P E 1967 Nearest neighbor pattern classification IEEE Trans. Inf. Theory 13 21–7
de Vries P C et al 2012 The impact of the ITER-like wall at JET on disruptions Plasma Phys. Control. Fusion 54 124032
de Vries P C et al 2014 The influence of an ITER-like wall on disruptions at JET Phys. Plasmas 21 056101
de Vries P C, Johnson M F, Alper B, Buratti P, Hender T C, Koslowski H R, Riccardo V and JET-EFDA Contributors 2011 Survey of disruption causes at JET Nucl. Fusion 51 53018
Duda R O, Hart P E and Stork D G 2000 Pattern Classification 2nd edn (New York: Wiley)
Gammerman A and Vovk V 2007 Hedging predictions in machine learning J. Mach. Learn. Res. 15 151–63
Hernandez J V et al 1996 Neural network prediction of some classes of tokamak disruption Nucl. Fusion 36 1009–17
Mahalanobis P C 1936 On the generalised distance in statistics Proc. Natl Inst. Sci. India 2 49–55
Murari A, Boutot P, Vega J, Gelfusa M, Moreno R, Verduolaeg G, de Vries P C and JET-EDFA Contributors 2013 Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions Nucl. Fusion 53 033006
Murari A, Vega J, Mazon D and Courregolengu T 2015 Preliminary numerical investigations of conformal predictors based on fuzzy logic classifiers Ann. Math. Artif. Intell. 74 155–78
Nene S A and Nayar S K 1997 A simple algorithm for nearest-neighbor search in high dimension IEEE Trans. Pattern Anal. Mach. Intell. 19 989–1003
Pautasso G et al 2002 On-line prediction and mitigation of disruption in ASDEX Upgrade Nucl. Fusion 42 100–08
Romanelli F et al 2012 Proc. 24th IAEA Fusion Energy Conf. 2012 (San Diego, USA)
Sengupta A and Ranjan P 2001 Prediction of density limit disruption boundaries from diagnostic signals using neural networks Nucl. Fusion 41 487–501
Shafer G and Vovk V 2008 A tutorial on conformal prediction J. Mach. Learn. Res. 9 371–421
Vega J, Dormido–Canto S, López J M, Murari A, Ramírez J M, Moreno R, Ruiz M, Alves D, Felton R and JET-EDFA Contributors 2013 Results of the JET real-time disruption predictor in the ITER-like wall campaigns Fusion Eng. Des. 88 1228–31
Vovk V, Gammerman A and Shafer G 2010 Algorithmic Learning in a Random World (New York: Springer)
Wroblewski D et al 1997 Tokamak disruption alarm based on neural network model of high-beta limit Nucl. Fusion 37 725–41
Yoshino R 2003 Neural-net disruption predictor in JT-60U Nucl. Fusion 43 1171–786