WPJET1-PR(17) 18198

A Murari et al.

# How to Investigate Disruption Physics using Machine Learning Tools

Preprint of Paper to be submitted for publication in Nuclear Fusion

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.

# How to Investigate Disruption Physics using Machine Learning Tools

by A.Murari[1,2], E.Peluso[2], M.Lungaroni[2], J.Vega[3] and M.Gelfusa[2] and JET Contributors*

*1) Consorzio RFX (CNR, ENEA, INFN, Universita' di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy.*

*2) Department of Industrial Engineering, University of Rome "Tor Vergata", via del Politecnico 1, Roma, Italy*

*3) Laboratorio Nacional de Fusión, CIEMAT. Av. Complutense 40. 28040 Madrid. Spain*

EUROfusion Consortium, JET, Culham Science Centre, Abingdon, OX14 3DB, UK

**Abstract**

In the last decades, lacking solid and detailed theoretical understanding, machine learning tools have been deployed in various Tokamaks to predict the occurrence of disruptions. Their results clearly outperform empirical descriptions of the plasma stability limits. On the other hand, all the machine learning techniques applied in practice show very poor "*physics fidelity*" (their mathematical models do not reflect the physics of the underlying phenomena) and limited interpretability. To overcome these limitations, in this paper a method is proposed to combine the predictive capability of machine learning tools with the formulation of the operational boundary in terms of traditional mathematical models, more suited to understanding the underlying physics. This is achieved by a novel combination of probabilistic Support Vector Machines and Symbolic Regression via Genetic Programming. The results are very positive. The obtained equations of the boundary between the safe and disruptive regions of the operational space classify with about 2.5 % of missed alarms and a similar percentage of false alarms. The models derived with the proposed data driven methodology therefore present better performance than traditional representations, such as the Hugill or the beta limit, by a factor. More importantly, they are compact and easy to grasp mathematical formulas, which are well suited to supporting theoretical understanding and benchmarking of empirical models. They can also help in setting up feedback schemes and can be deployed efficiently in real time.

*See the author list of "X. Litaudon et al 2017 Nucl. Fusion 57 102001

Keywords: Disruptions, Prediction, Probabilistic SVM, Symbolic Regression, Genetic Programming, Data Driven Theory, Knowledge Discovery

Corresponding author: emmanuele.peluso@ uniroma2.it

## 1 Operation-based description of disruptions in Tokamaks

Many natural and man-made systems can look very resilient but in reality are prone to catastrophic collapse. Some of these collapses are quite straightforward to interpret and do not seem worthy of particular attention because, given the proper precautions, they are relatively easy to avoid. Others are very subtle and extremely difficult to predict. Earthquakes, and in general failures due to atmospheric phenomena, belong to this category. In the last years, various efforts have been devoted to develop mathematical tools more appropriate to investigate and predict these catastrophic and typically rare events. Machine learning tools, of the type described in this paper, constitute an additional family in the arsenal of mathematical approaches which can be used to study catastrophic events [1].

In the field of Magnetic Confinement Nuclear Fusion, disruptions are the most striking example of catastrophic failures difficult to predict. Therefore they are one of the most severe problems to be faced by the Tokamak configuration in the attempt to design and operate commercial reactors. Disruptions are also a not negligible issue for the present largest devices, such as JET, and indeed they pose a not irrelevant constraint to high current operation. The percent of disruptions allowed in ITER is quite limited. Certainly, JET disruptivity with the new ITER Like Wall (ILW) remains too high for the next generation of devices, since at high current the percentage of disruptions can exceed 30%. In DEMO even one unmitigated disruption could severely damage the reactor [2].

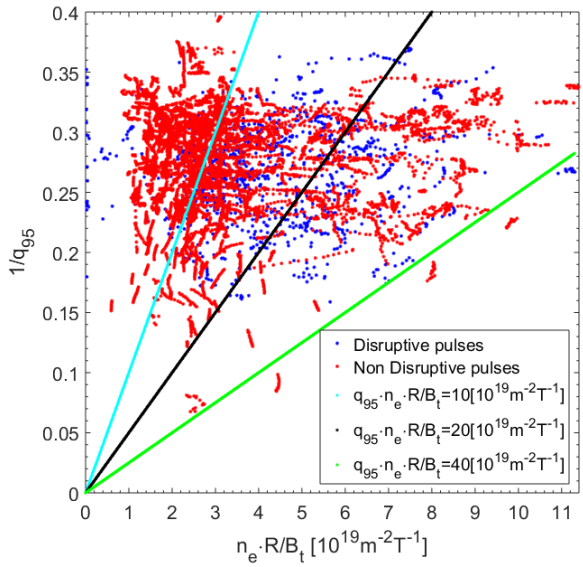


**Figure 1. Hugill plot for a large database of JET with the ILW covering campaigns C29-C31. The disruptive and safe discharges overlap completely in this space.**

Since they constitute a potential serious hazard to the integrity of Tokamak devices, disruptions are the subject of extensive studies at present. From the perspective of ultimate remedial actions which can be undertaken, various methods of mitigation are being investigated, particularly massive gas injection and shatter pellets. The main objective of massive gas injection consists of limiting the energy conducted directly to the wall by converting it into radiation. On the other hand, this conversion method can pose other hazards

to the machines, such as the generation of runaway electrons, and shatter pellets are indeed aimed exactly at extinguishing the beams of such fast particles. A complementary approach to manage the problem of disruptions is based on avoidance, i.e. on the sufficiently advanced detection of problems in the discharges and the consequent implementation of remedial actions to avoid the abrupt termination of the plasma.

From an operational perspective, robust and reliable prediction methods are a prerequisite to any mitigation or avoidance action. Unfortunately, the theoretical understanding of the causes of disruptions is not sufficient to guarantee reliable predictions, particularly on the time scales required for avoidance. Lacking solid and detailed theoretical understanding of disruptions, it has been attempted to develop an operation-based description of Tokamak plasmas, aimed at determining the boundaries of the safe space in terms of physically controllable quantities. One of such empirical descriptions of plasma stability is the so called Hugill diagram, which combines the low q and density limits [3]. The low q limit is expressed in terms of $1/q_{95}$, where $q_{95}$ is the safety factor at 95 % of the plasma radius. The density limit is typically expressed in terms of the Murakami factor $n_eR/B_T$ where $n_e$ is the mean electron density, R the plasma major radius and $B_T$ the toroidal field. For a large JET database with the ILW (see Section 6.2), the Hugill diagram is reported in Figure 1. Unfortunately such a plot has very poor predictive and interpretative capability, since the disruptive and non disruptive examples overlap almost completely and there is practically no frontier between the safe and disruptive regions of the operational space.

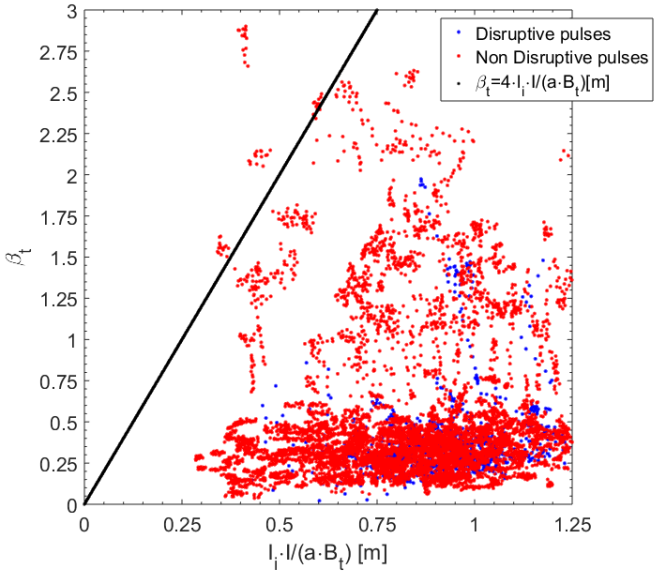


**Figure 2. The beta limit plot for a large database of JET with the ILW covering campaigns C29-C31. Also in this space the disruptive and safe discharges overlap completely.**

Similar considerations apply to another popular diagram, used to investigate the so called beta limit. The parameter β is typically used to quantify the level of the plasma pressure compared to the magnetic pressure. It is therefore natural to expect that pressure

driven instabilities could limit the level of β achievable in a certain configuration. This limit is typically represented as a function of the parameter $I*l_i/(aB_T)$, where I is the plasma current, $l_i$ the internal inductance and $a$ the minor radius. The β-limit plot for various campaigns of JET with the ILW is reported in Figure 2. Again inspection of the plot reveals that, in this space, it is practically impossible to separate the disruptive from the safe operational regions. Therefore, from a practical point of view, these representations have poor predictive capability and cannot be used for any form of forecasting. Also from the interpretation point a view they leave a lot to be desired, since they do not provide clear empirical evidence about the real basis for the difference between the safe and disruptive regions of the operational space. .

The inadequacies of theoretical and empirical models of disruptions have motivated the development of data driven predictors. In this perspective, various machine learning methods have been developed. They range from neural networks to Fuzzy logic classifiers [4-12]. In the last years, a new classifier, APODIS, based on Support Vector Machines (SVM) has been deployed in JET real time network and has provided very satisfactory performance in terms of both success rate and false alarms, in a long series of campaigns without any need for retraining. Manifold learning tools, such as Self Organising Maps and Generative Topographic Maps, and simple classifiers based Geodesic distance on Gaussian manifolds have provided very good results also in terms of automatically determining the disruption type many tens of ms in advance of the beginning of the current quench [13-15].Even if these data driven tools are providing quite impressive results, their main problem, particularly in the perspective of the next step devices, is the amount of examples required for training. In large machines such as ITER, it would be practically impossible to collect hundreds of examples to train the most performing machine learning tools such as APODIS. In the last couple of years therefore many efforts have been devoted to the developments of parsimonious data driven techniques, which can provide good success rate of prediction after a few tens of disruptions and even after the first disruption[16,18].

The main remaining issues with these advanced machine learning tools are now their *physics fidelity* and the interpretability of their results. They have shown the potential to learn very efficiently from the provided examples but they are formulated in such a way that does not necessarily reflect the dynamics behind the phenomenon. Since the resulting models are also very difficult to interpret, in the present format machine learning tools cannot really contribute to the interpretation of the physics behind disruptions. This aspect is quite

worrying given the fact that they have therefore to be considered black boxes, whose extrapolation to larger devices could be questioned.

To increase the contribution of machine learning tools to the interpretation of the physics, a new methodology has been developed to profit from the knowledge acquired by the machine learning tools but presenting it in a more traditional format, in terms of manageable formulas, which can be used both as a guide for analysis developments and as a benchmark of theoretical models. This approach reconciles the prediction and knowledge discovery capability of machine learning tools with the need to formulate the results in such a way that they can be related to physical theories capable of extrapolation to larger devices.

The main technique, to derive physically meaningful models from machine learning tools, consists of the following steps:

1- Training the machine learning tools for classification, i.e. to discriminate between disruptive and non-disruptive examples

2- Determining a sufficient number of points on the boundary between safe and disruptive regions of the operational space identified by the machine learning tools

3- Deploying Symbolic Regression via Genetic Programming to express the equation of the boundary from the points previously obtained in a physically meaningful form

The potential applications of the proposed new methodology are many. Two important cases will be discussed in detail in the following. The first is the data driven derivation of the equation of the boundary between disruptive and non-disruptive regions of the operational space, obtained without any "*a priori*" assumption on the form of the models. Such an example is meant to illustrate the exploratory power of the developed techniques. The second main application relates to the complementary problem of building models on the basis of constraints derived from previous works or theoretical considerations.

In the proposed approach, the knowledge discovery step is based on Support Vector Machines (SVM), whose mathematical background is summarised in the next Section. This choice is driven by the properties of structural stability of SVMs, which guarantee very high success rates provided the training set is adequate. In addition to traditional Support Vector Machines, the proposed method is also adapted to a probabilistic version of SVM, to show in general how the approach can be particularised for machine learning tools, which provide

outputs of different nature. To formulate the output of SVM in a physically realistic and interpretable way, extensive use is made of Symbolic Regression via Genetic programming; these tools are therefore described in Section 3.

The actual combination of the two methods, to provide the equation of the boundary between two regions of the operational space, is described in detail in Section 4 and some synthetic examples are reported in Section 5. JET database with the ITER Like wall is introduced in Section 6 and the results, in terms of describing the boundary between the disrupting and non disruptive regions of the operational space, are the subject of Sections 7 and 8. Discussions and lines of future interpretation are the subject of the following Section 9.

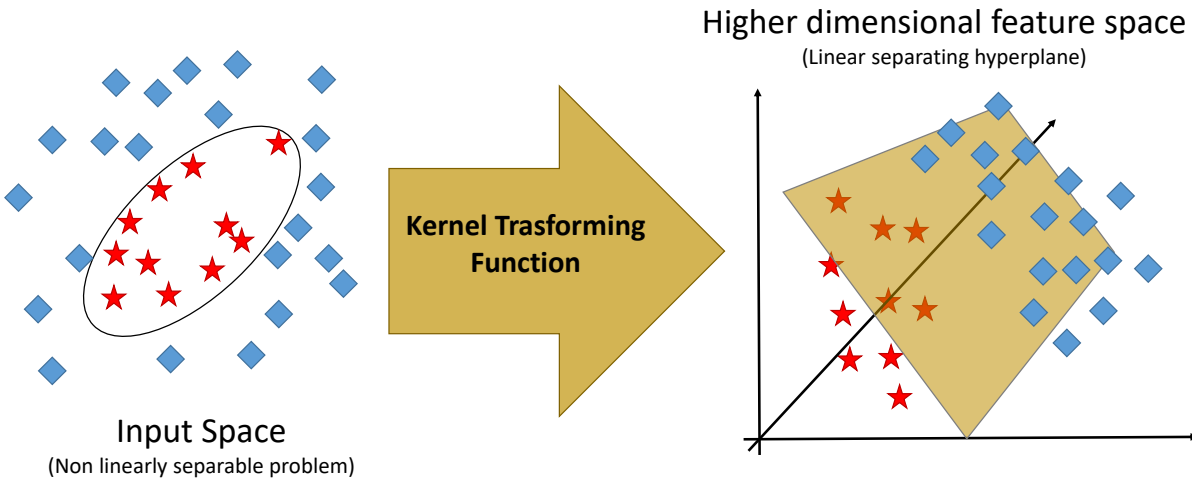


**Figure 3. The basic principle of the SVM approach to classification: the projection to a higher dimensional space with a kernel to find the best separating hyperplane.**

## 2 Introduction to SVM for Classification

Machine learning tools are typically deployed to derive information directly from the data in two main eventualities: a) in *theory-less* applications, such as many in the private sector, when there is no ambition to formulate a physically meaningful mathematical model b) when the problems to address are so complex that is very difficult or even impossible to derive theoretical models from first principles. The main objective of the work presented in this paper relates to the second case and consists of defining a methodology, which can allow obtain satisfactory mathematical models from the results of machine learning tools instead of from basic principles.

*2.1Traditional SVM*

SVMs are very powerful machine learning tools, with various very desirable properties for scientific applications [19]. They are used as classifiers for the studies described in this paper. In intuitive terms, given a set of input examples, which belong to two different classes, the SVM maps the inputs into a high-dimensional space through some suitable non-linear mapping. In this high dimensional feature space, an optimal separating hyperplane is constructed in order to minimize the risk of misclassification. The minimization of the error risk is obtained by maximizing the margins between the hyperplane and the closets points, the support vectors, of each class. This is achieved by a careful selection of the constraints of a suitable functional to maximize. The hyperplane is expressed in terms of a subset of points of the two classes, named *Support Vectors* (SV). The main idea behind the SVM approach is illustrated in Figure 3.

Once the support vectors have been determined, the SVM boundary between the two classes can be expressed in the form

$$d(\boldsymbol{x}) = \sum_{i=1}^{\ell} \alpha_i y_i H(\boldsymbol{x}_i, \boldsymbol{x}) \qquad (1)$$

where $d(\boldsymbol{x})$ is the distance from the input $\boldsymbol{x}$ to the hyper-plane that separates the two classes and, hence, the hyper-plane points satisfy $d(\boldsymbol{x})=0$.

The rule to classify a feature vector $\mathbf{u}$ as disruptive (class $C_{Dis}$) or non-disurptive (class $C_{Safe}$) is given by:

$if\ sgn\big(D(\boldsymbol{u})\big) \geq 0$

$\boldsymbol{u} \in C_{Disr}$

$otherwise$

$\boldsymbol{u} \in C_{Safe}$

where sgn(t) is the sign function.

Given the structural stability that they achieve by implementing the margins, SVMs are very powerful tools and have performed extremely well in the case of disruption prediction, as demonstrated in real time by APODIS. Their hyperplane can therefore be considered a very good approximation of the boundary between the disruptive and not disruptive regions of the operational space. On the other hand, their mathematical representation of the boundary is of the type of equation (1). Relations with the mathematical

structure of equation (1) bear no resemblance to the underlying dynamics of the physical phenomenon under investigation. Moreover equation (1) is extremely non intuitive. In the case of disruptions on JET, the equation of the hyperplane can easily comprise more than 500 support vectors and therefore the equation of the hyperplane contains an equal number of addends. To obtain an equation for the hyperplane, which better reflects the physics of the phenomenon and easier to interpret, the method of Symbolic Regression via Genetic Programming has been adopted, as explained in the next Section. First an overview of probabilistic SVM is provided, since this is the tool mainly used in the rest of the paper for the exploratory phase of the analysis.

*2.2 Probabilistic SVM*

The availability of classifiers, which can output a probability, would be extremely useful in most applications. Unfortunately, traditional SVM provide only a distance to a hyperplane, in the from reported in equation (1). Their basic version has therefore to be extended to associate a probability to the outputs of their classification [20-22].One possible solution consists reformulating the SVM output in terms of a probability with the Bayes rule according to the formula:

$$P(y=1|D) = \frac{p(D|y=1)P(y=1)}{\sum_{i=-1,1} p(D|y=i)P(y=i)} \qquad (2)$$

In equation (2) D are the data and y indicates the label of one of the classes (the disruptive one for example to fix the ideas). P(y=1) is the prior probability of disruption and p(D|y=1) is the likelihood, the probability of the data given the fact that the time slice in question is disruptive. Therefore, to convert the outputs of traditional SVM to probabilities, two quantities have to be determined: the prior probability and the likelihood. In our application, the natural choice of the prior probability is the percentage of time slices seen so far in the campaign for the class to which the SVM classifies the new example. The most challenging aspect of relation (2) resides in the evaluation of the likelihood. A solid and reliable estimate of the likelihood would require much more data than that available. Moreover, in any case probability density estimation is always a delicate and time consuming process. Theoretical investigations and practical considerations have shown that, for our application, one alternative advantageous solution consists of remapping the distance to the hyperplane to a probability by using a sigmoid function [20,21]:

$$P(y = 1|d) = \frac{1}{1+exp(Ad+B)} \qquad (3)$$

In equation (3) A and B are two fitting parameters, whereas *d* is the distance of the examples to the SVM hyperplane. Equation (3) therefore allows to convert directly the distance to the hyperplane, provided by traditional SVM, into a probability. This conversion takes place after the training; the distances of the examples in the training set are used to fit the parameters of the sigmoid (3). The sigmoid is constrained to be centred on the hyperplane, because points at distance zero from it have equal probability of belonging to any of the two classes.

## 3 Symbolic Regression via Genetic Programming for interpretability

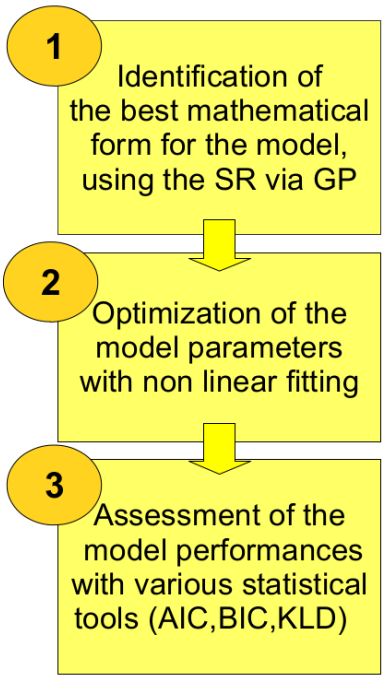


**Figure 4. Overview of the main steps of the proposed methodology used to identify the best mathematical models directly from the data.**

As mentioned in the previous section, this paper describes the application of Symbolic Regression (SR) via Genetic Programming (GP) to identifying the boundary between safe a disruptive regions of the operational space. The main advantages of the proposed approach consist of: a) practically eliminating any assumption about the mathematical form of the boundary b) allowing to express the equations of the boundary in mathematical forms, which match the physics of the disruption and permit comparison with theory. The methods developed indeed allow identifying the most appropriate mathematical expressions for the boundary equations and therefore have the potential to better interpret the experimental data for the boundary of the operational space.

The method consists of testing various mathematical expressions to fit a given database. The main steps to perform such a task are reported in Figure 4. First of all, the various candidate formulas are expressed as trees, composed of functions and terminal nodes. The function nodes can be standard arithmetic operations and/or any mathematical functions, squashing terms as well as user-defined operators [23,24]. Terminal nodes are typically physically measurable quantities. This representation of the formulas allows an easy implementation of the next step, symbolic regression with Genetic Programming. Genetic Programs (GPs) are computational methods able to solve complex optimization problems [23,24]. They have been inspired by the genetic processes of living organisms. They work with a population of individuals, e.g mathematical expressions in our case. Each individual represents a possible solution, a potential equation for the boundary between the safe and disruptive regions of the operational space in the application presented in this paper. A fitness function (FF) is used to measure how good an individual is with respect to the database. The FF is basically a metric determining how good an individual is in solving the problem at hand given the database. A higher probability to have descendants is assigned to those individuals with better FF. Therefore, the better the adaptation (the better the value of the FF) of an individual to a problem, the higher is the probability that its genes are passed to its descendants.

In practice, the first step of the method is the generation of the initial population of formulas for the operational boundaries and then the algorithm finds how well an element of the population works, assessing its performance with the FF. In the second phase, as with most evolutionary algorithms, genetic operators (Reproduction, Crossover and Mutation) are applied to individuals that are probabilistically selected on the basis of the FF, in order to generate the new population. That is, better individuals are more likely to have more descendants than inferior individuals. When a stable and acceptable solution, in terms of

**Table I: Types of function nodes included in the symbolic regression used to derive the results presented in this paper, $x_i$ and $x_j$ are the generic independent variables.**

| **Function and operator class** | **List** |
|---|---|
| **Arithmetic operators** | constants,+,-,*,/ |
| **Exponential functions** | exp($x_i$),log($x_i$),power($x_i$, $x_j$), power($x_i$,c) |
| **Squashing functions** | logistic($x_i$),step($x_i$),sign($x_i$),gauss($x_i$),tanh($x_i$), erf($x_i$),erfc($x_i$) |

complexity, is found or some other stopping condition is met (e.g., a maximum number of generations or acceptable error limits are reached), the algorithm provides the solution with best performance in terms of the FF.

In this work, the models are composed of functions and terminal nodes and can be represented as a combination of syntax trees. The function nodes included in the analysis performed in this paper are reported in Table I. It is worth emphasising that no detailed hypothesis on the mathematical structure of the final equation is assumed "a priori". SR via GP extracts the most suitable formulas on the basis of the input data. On the other hand, constraints on the final results can be implemented by selecting the most appropriate basis functions or by constraining the structure of the trees. So the proposed approach is data driven and does not force the solution to belong to a specific class of models; at the same time it can impose a mathematical formulation more appropriate to the phenomenon at hand.

In addition to the basis functions, the fitness function is the other crucial element of the genetic programming approach and it can be implemented in many ways. To derive the results presented in this paper, the AIC criterion (Akaike Information Criterion) has been adopted [25] for the FF. The AIC form used is:

$$AIC = 2k + n \cdot \ln(RMSE) \qquad (4)$$

In equation (7), RMSE is the Root Mean Square Error, $k$ is the number of nodes used for the model and $n$ the number of $y_{\text{data}}$ provided, so the number of entries in the database (DB). The FF parameterized above allows considering the goodness of the models, thanks to the RMSE, and at the same time their complexity is penalised by the dependence on the number of nodes. The parameters of the mode obtained with SR via GP are typically refined with appropriate nonlinear fitting routines. In addition to improving their values, this step allows associating confidence intervals to the parameters of the models.

To assess the quality of the final models the well-known criteria of BIC (Bayesian Information Criterion) and Kullback-Leibler (KLD) divergence have been used. The BIC criterion is defined as:

$$BIC = n \cdot \ln\left(\sigma^2_{(\epsilon)}\right) + k \cdot \ln(n) \qquad (5)$$

where$\epsilon = y_{\text{data}} - y_{\text{model}}$ are the residuals, $\sigma^2_{(\epsilon)}$ their variance and the others symbols are defined in analogy with the AIC expression. Again the better the model, the lower its BIC.

Then the aim of the KLD is to quantify the difference between the computed probability density functions, in other words to quantify the information lost when $p\left(\vec{y}_{model}(\vec{x})\right)$ is used to approximate $q\left(\vec{y}_{data}(\vec{x})\right)$ [27]. The KLD is defined as:

$$KLD(P\|Q) = \int p(x) \cdot ln\big(p(x)/q(x)\big)dx \quad (6)$$

Where the symbols have been defined as above. The Kullback-Leibler Divergence assumes positive values and is zero only when the two probability distribution functions (pdfs), *p* and *q*, are exactly the same. Therefore the smaller the KLD is, the better the model approximates the data, i.e. the less information is lost by representing the data with the model.

It is worth mentioning that in traditional applications of SR via GP the method has been used to perform actual regression and therefore to identify functions [27-29]. The application to identification of the boundary between different regions of the operational space is a new application in fusion, reported for the first time in this paper. Mathematically this problem is more involved; indeed, in general, the boundaries between disruptive and safe region of the operational space or between different disruption types do not need to be functions. On the other hand, the methodology of SR via GP is much more general and it is not limited to function. It has indeed been also verified numerically that, provided the right choice of basis functions is chosen, the approach can identify also closed surfaces.

## 4 Combining SVM and Symbolic Regression to Identify the Equation of the Boundary

This section introduces the details of the mathematical procedure to obtain the equation of the boundary between disruptive and non-disruptive operational regions, by applying symbolic regression to the output of SVM classifiers. The method consists mainly of two parts. First points on the SVM separating hyperplane are generated and then they are fitted with symbolic regression. In the case of probabilistic SVM the first task is banal. Indeed the available tools allow plotting iso-probability surfaces. It is therefore sufficient to extract the points which correspond to the chosen level of probability. These points can then

be fitted with SR via GP to obtain the required equation. The situation is a bit more involved in the case of traditional SVM, which provide a distance to the separating hyperplane and not directly points on the hyperplane. The solution for this more complicated case is reported in the following for completeness sake and for the benefit of systems using the traditional version of the SVM.

In order to interpret the results produced by the traditional SVM, as already mentioned, the first step consists of determining a sufficient number of points on the hypersurface separating the two classes. These points can be then given as inputs to the SR to obtain a more manageable equation for the hypersurface. To obtain the SVM hypersurface points, a mesh is built first, with resolution equal or better than the error bars of the measurements used as inputs to the SVM. In this step, a refined mesh throughout the domain defined by the ranges of variables is generated; therefore, if the problem presents $n$ dimensions and $m$ grid points are generated for each dimension, the grid will consist of $m^n$ number of grid points. Obviously, more grid points and a better refined mesh lead to more accurate results; therefore, the total number of grid points can be set based on computational

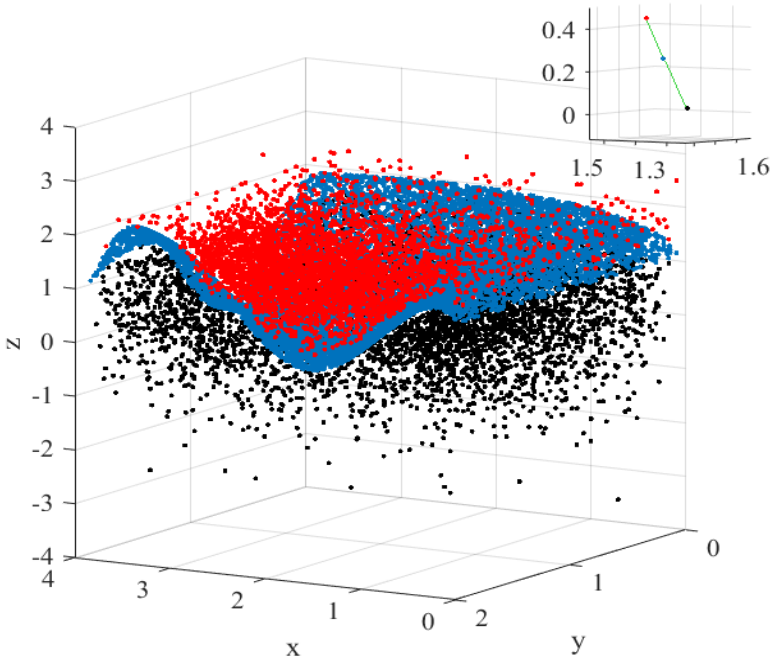


**Figure 5: SVM hypersurface points for synthetic, linearly separable data set. The black and red points belong to the two classes. The blue points are the ones lying of the hypersurface and obtained with the method explained in the text and illustrated pictorially in the insert.**

limitations. On the other hand, suitable criteria are available for selecting more efficiently the number of intervals in different directions. The first one consists of allocating more intervals along the direction of stronger curvature. The second alternative is that of allocating a higher number of mesh points in the direction of the dependent variable, to be sure that the points will be selected for the hyper-surface are close enough to the real hyper-plane.

After building the grid, the algorithm starts selecting the Support Vectors (SVs) on the positive side of the hypersurface and moves towards the SVs on the other side, one point on the mesh at the time. At each step, the distance to the hypersurface is computed using the already trained SVM. If the distance remains positive, the process is repeated since the new point remains on the same side of the hypersurface. When the distance of a new point changes sign, the two points with different signs are on opposite sides of the hypersurface separating the two classes. They can therefore be considered points on the hypersurface, since, by construction of the mesh, these points, for which the distance changes sign, are within a distance from the hypersurface equal or smaller than the error bar of the features (typically measurements). Therefore, for all practical purposes, the points found as previously described are sufficiently close to the hypersurface to be considered on it. This way to obtain SVM hypersurface points for synthetic data is shown pictorially in Figure 5. The support vectors on either side of the hypersurface are given a different colour and a line connecting the two sides of the hypersurface is drawn in the insert.

Once it has been verified that a sufficient number of points close to the hypersurface have been found, the equation of the hypersurface itself can be estimated using SR via GP. Indeed the points, identified with the procedure just described, are on the boundary between the two classes. Therefore the equation of that surface is the equation of the boundary between the two classes. The quality of the obtained equation can be assessed first with the statistical indicators described in Section 3. Another very important step to prove the quality of the obtained equations consists of testing their success rate of classification, for the same examples used for training the SVM.

## 5 Numerical tests and results

The procedure described in the previous section has been subjected to a systematic series of numerical tests. The results have always been positive and the proposed technique has always allowed recovering the original equations describing the boundary between the

two classes. In the following the detailed procedures for these numerical tests are described in detail and some results presented.

*5.1 Overall procedure for producing synthetic data*

The main technique to produce synthetic data to test the methodology consists of the following 6 steps:

1- Definition of an initial function for the boundary

2- Generating samples of the two classes from the function

3- Training the SVM for classification

4- Building an appropriate mesh on the domain

5- Determining a sufficient number of points on the hyper-surface identified by the SVM

6- Deploying symbolic regression to obtain the equation of the hypersurface from the points previously generated

In the case of probabilistic SVM, point 4 and 5 collapse to a single very easy step, since it is possible to directly obtain the points at the required level of probability from the machine learning tool.

In the rest of this subsection, more details about this procedure are provided. To fix the ideas, the discussion is particularised for the case of traditional SVM. In the first step, an initial function as a combination of arithmetic, trigonometric, and exponential operators of independent variables $x_i$ is defined. In general, this function can be written as follows:

$$y = f(x_1, x_2 \ldots) \qquad a_1 < x_1 < b_1 \quad a_2 < x_2 < b_2 \quad etc$$

In the second step, a sufficient number of random points is generated in the relevant range of the variables and from them the dependent variable y is calculated. Then, a positive offset and some random values are added to the y for half of the data to produce the first class; a negative offset and some random values are added to y for the other half to produce the second class. Adding random values is meant to simulate the effect of the noise. The statistics of this additive component can therefore be adapted to the experimental measurements available; in our case the noise is assumed to have a Gaussian distribution. The equations for producing the two classes can be summarized as follow:

$y_1 = y$ + *random data between 0 and L + offset* $\quad$ *L= bulk thickness of the data*

$y_2 = y$ - *random data between 0 and L - offset* $\quad$ *L= bulk thickness of the data*

where $y_1$ and $y_2$ are the values for the first and second class, respectively.

In the third step, an SVM is trained. The method used to find the separating hyperplane is "Sequential Minimal Optimization" [19]. Depending on the level of random noise, different success rates can be obtained. For the numerical tests presented in the following, the success rate in the classification of the SVM is always very close to 100%.

**Table II: General GP parameters for the calculations of the boundary equations**

| GP Parameters | Value(s) |
|---|---|
| **Population size** | 500 |
| **Selection method** | Ranking and Tournament |
| **Fitness function** | Gaussian distribution |
| **Constant range** | Integers between -10 and 10 |
| **Maximum depth of trees** | 5 |
| **Genetic operators (Probability)** | Crossover (45 %)<br>Mutation (45 %)<br>Reproduction (10 %) |

In the fourth step, a mesh on the domain is built in order to identify points sufficiently close to the hypersurface.

The fifth step consists of the identification of the points sufficiently close to the hypersurface, with the algorithm described in Section 5.

In the sixth step, the selected hypersurface points are used as inputs to the symbolic regression code, to find the appropriate formula for describing the hypersurface. The settings adopted to run the GP implementing the SR are presented in Table II:

*5.2 Example of two independent variables*

In this section, an example is provided to illustrate the applicability and capability of the presented methodology. As a representative test, a function comprising trigonometric and arithmetic operators has been defined. The function and ranges of the variables are:

$$y = \sin(x_1) + x_2 \qquad -3 < x_1 < 3 \qquad -2 < x_1 < 2 \qquad (7)$$

For this example, each dimension of the domain has been subdivided in one hundred intervals, producing one million mesh points ( $100^3$ ). After carrying out the six-step procedure previously described, the following expression has been obtained:

$$y = 0.985 \, ( \sin(x_1) + x_2 ) \qquad (8)$$

SR via GP converges on a final expression that is in excellent agreement with the initial function describing the boundary between the two classes, even without making recourse to the non-linear fitting step. Figure 6 presents the results of this example in pictorial form.

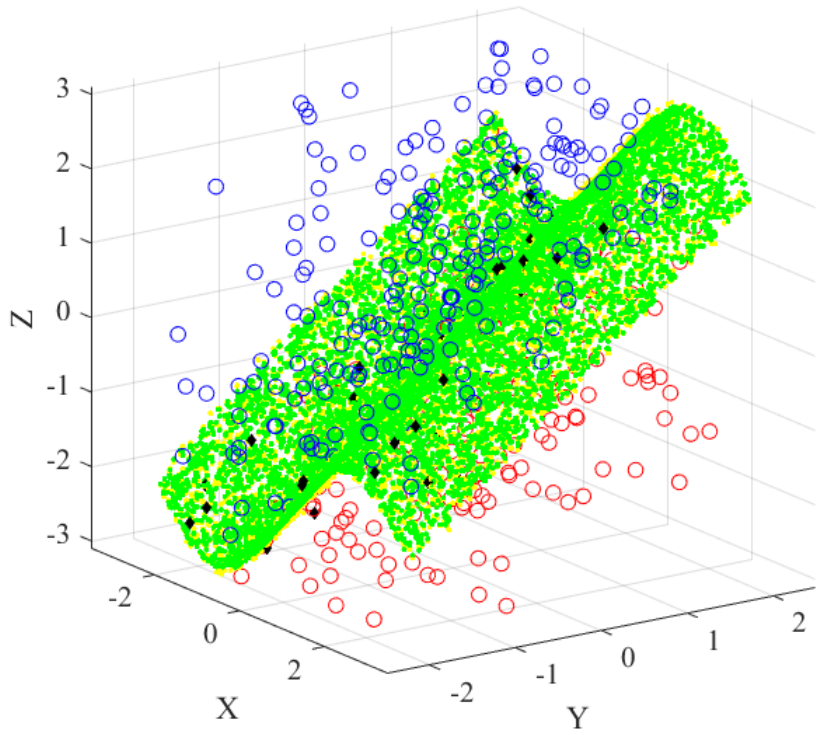


**Figure 6: Points and surfaces of the example in subsection 5.2. Green are points generated from the initial function, blue points are the points belonging to the first class, Red points are the points belonging to the second class, black diamond identify support vectors of the first class and the Yellow surface identifies the hyper-surface obtained with the SR via GP.**

*5.3 Example for four independent variables*

As mentioned, there is no conceptual difficulty in applying the proposed methodology to higher dimensional problems. Of course, the computational resources required increase exponentially with the number of independent variables (the so called curse of dimensionality). Also the quality of the measurements must be adequate and the number of examples sufficient. But these are problems related to the available computational power and/or the quality of the data; in no way they affect the applicability of the proposed technique. Indeed it has been verified with a series of systematic tests that, with adequate level of computer time, problems in higher dimensions can also be solved.

In this subsection, we describe the results of the application of the SVM-GP methodology to a more complex and noisy database. This example can be considered of the level of complexity of the actual problem discussed in the following sections, namely the determination of the disruptive region in terms of traditional signals available in real time. To this end, a five-dimensional synthetic database has been generated with the characteristics described in Table III.

**Table III Settings for testing SVM-GP on a five-dimensional synthetic database**

| **Steps:** | **Values:** |
|---|---|
| Initial Function | $y = sin( x_1 + x_2 ) - 0.5\, x_3\, x_4$ |
| Ranges of Variables | $-1.5 < x_1 < 1.5$ & $-2 < x_2 < 2$<br>$0 < x_3 < 2$ & $2 < x_4 < 4$ |
| Number of Nodes for Each Class | *2000* |
| Thickness of the data's bulk | *3* |
| Offset | *10% of y domain* |
| Classification Noise | *~ 4%* |

Then the procedure of finding the best sigma for the SVM has been applied and the best sigma for the classification is equal to 0.6. The final accuracies of classification for the train and test data are presented in Table IV.

**Table IV: The accuracies obtained by the SVM for the train and test data on the classification of the synthetic database with the best sigma that equals to 0.6**

| Database Type: | Classification Accuracy in Percent: |
|---|---|
| Train Data | 96.1337 |
| Test data | 96.0422 |

After generating the grid and finding the hyper-surface points, SR via GP has been applied and the following expression for the hyper-surface has been obtained:

$$y = 0.9334 \sin (0.9190 ( x_1 + x_2 ) ) - 0.5010 x_3 x_4 \quad (9)$$

The obtained equation is in good agreement with the initial function, reported in Table III. The quality of this estimate can be confirmed by comparing the success rate of the SVM and of the equation found by SR via GP. The classification success rate of the equation found with SR is reported in Table V (to be compared with the results reported in Table IV).

**Table V: The accuracies obtained for the train and test data for the classification of the synthetic database with the expression obtained via GP**

| Database Type: | Classification Accuracy in Percent: |
|---|---|
| Train Data | 96.1060 |
| Test data | 96.3061 |

From the comparison of the success rates obtained via SVM and with the derived mathematical equations, it can be concluded that the SVM-GP method has excellent performance, even for more complex databases and in higher dimensionality, for interpreting the SVM hyper-plane as a hyper-surface equation.

*5.4 Computational Requirements*

As an indication about the computational resources required for the application of the proposed technique, the run time for the example of 5 variables has been calculated. Using a computer with 8 cores and 24 gigabyte of RAM (an Intel Xeon E5520, 2.27 GHz, 2 processors), with Windows 64 bit operating system, finding the hyper-surface points takes 3 hours and the GP calculation 48 hours. The number of points on the grid is $16^4 * 51$ is; 16 for the four independent variables and 51 for the dependent one. The run time to train the SVM is

not a major issue since it is typically of an order of magnitude shorter than the GP calculations and therefore negligible compared to the other steps of the procedure.

## 6 Database of JET with the ILW wall

In building the database, the intentional disruptions have been excluded from the training. Only time slices, whose plasma current exceeds 750 kA, have been considered but no other general selection has been implemented. All the signals have been resampled at 1kH frequency. Alarms, which are launched 10 ms or less from the beginning of the current quench, are considered tardy, since 10 ms is the minimum time required on JET to undertake mitigation action. Alarms triggered more than 2.5 s before the beginning of the current quench are considered early.

In more detail, the campaigns C29 to C31 have been considered. After proper

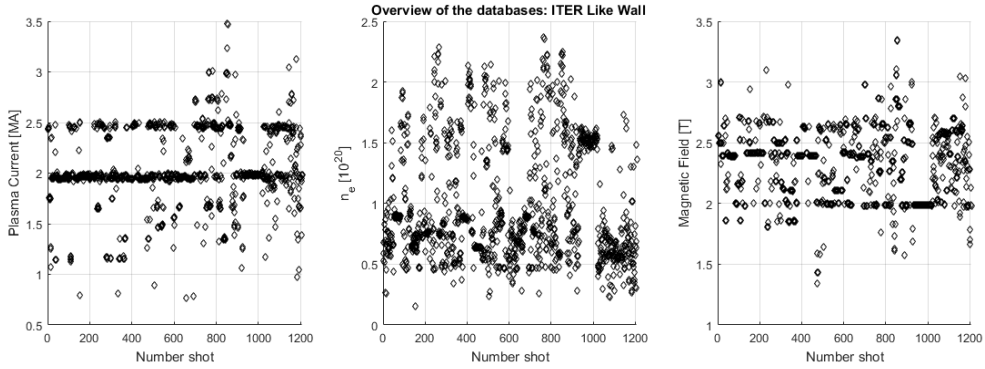


**Figure 7 .Overview of the database**

cleaning and validation of the DB, overall 187 disruptive and 1020 non disruptive shots are included, unless differently specified (in some analyses at the end of the paper the need to consider additional variables not always available will require to slightly reduce the statistical basis). A plot showing the operational space covered by the database is shown in Figure 7. As can be seen from the reported plots, the set of discharges considered is representative of JET operation with the ILW; therefore the same quality of the results are to be expected also if the methodology is applied to other campaigns.

## 7 A Data driven Model for JET with the ILW

To illustrate the potential of the proposed methodology, in this section the task of extracting a mathematical model directly from the data is assessed. The case discussed in this section is an example for which the "a priori" knowledge about the problem is kept to a minimum. The tools are applied to the database without any bias. The objective of this example is to show how the technique proposed can be used to obtain a practical, easily interpretable and implementable formula, without necessarily providing a completely satisfying physical model. To expedite the formulation of the models, a specific training method is adopted, as described in the next subsection. This approach allows reducing to a minimum the number of examples required for the training, which is extremely important for exploratory applications, since the computational time required by the SVM increases exponentially with the number of examples. In subsection 7.2 the method is applied to a database of JET with the ILW.

*7.1 Adaptive approach to efficient training*

In line with previous cases [16,17], a quite simple approach has been implemented for the training. The predictors needs at least one disruptive and one non disruptive case to build the first model. In the campaigns analysed, the first disruption occurred after a while and therefore the first model was obtained after the first disruption. For the disruptive discharge, 12 ms before the beginning of the current quench have been divided in 4 intervals of 3 ms each and the averages of these three intervals have been used as input to the training. The 10 discharges prior to the first non disruptive have been used as examples for the safe case. For each of these discharges, a random interval of 40 ms, with plasma current above 750 kA, has been divided in four 10 ms ranges and the averages over those subintervals have been used as inputs for the training.

The model derived as previously described has been used for the following discharges until the first misclassification. When the previous model misses a disruption or causes a false alarm, the shot not properly classified is included in the training set. In this way a new model is determined, which is deployed to analyse the following discharges until the next error, which provides an example for a new retraining. For every retraining, if the previous error is a missed alarm, again the same information about this shot is included in the training set (12 ms before the beginning of the current quench are divided in 4 intervals of 3 ms each and the averages of these three intervals are the additional features). If the error requiring the retraining is a false alarm, an interval of 40 ms before the alarm is divided in four 10 ms ranges and the averages over those subintervals are the new features. In the case of the false

alarms a longer interval has proved better for the predictor to recognise that the discharge is in a safe region of the operational space.

It is worth pointing out that the adopted procedure for the training of the probabilistic SVM is very efficient. Only the most relevant information is retained in the training set. Therefore, the computational requirements of the SVM training are kept to a minimum. The version of the adaptive training adopted in this paper has been devised to maximize the success rate of the classification, in order to generate the best mathematical models. A version compatible with real time applications has been already presented in [18]. For the specific database analysed in this paper, adopting the real time compatible training methodology would not cause any significant reduction in performance and would not alter the conclusions in any noticeable way.

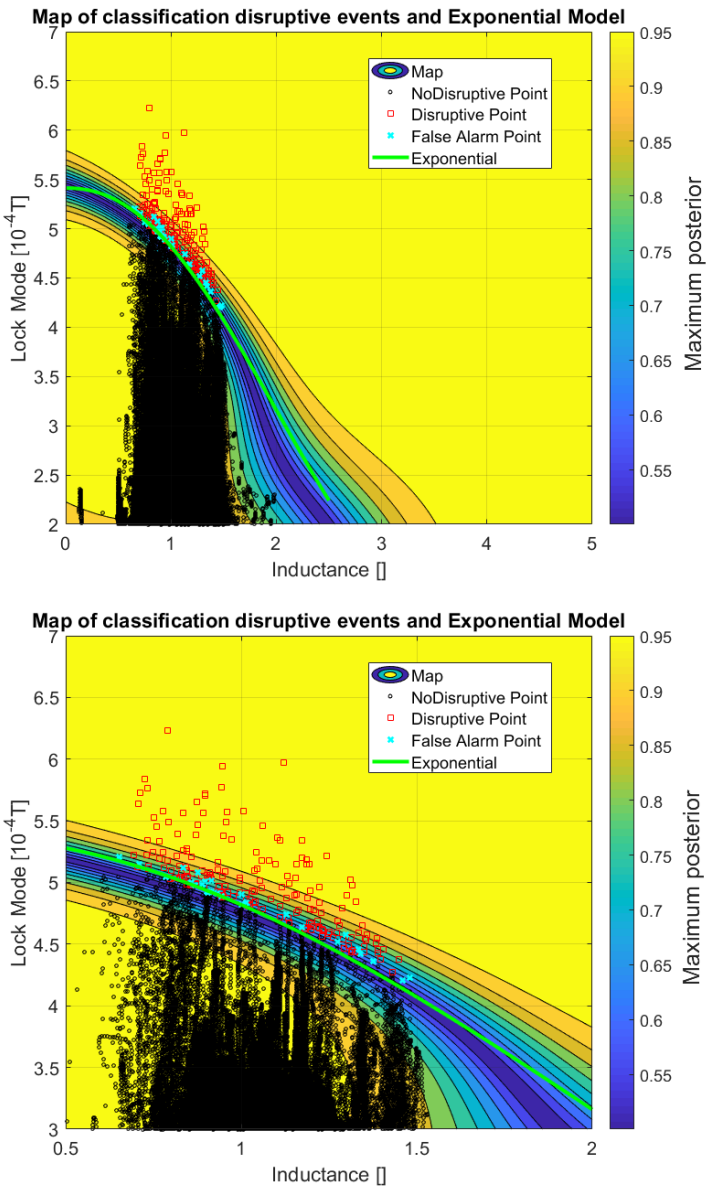


**Figure 8. Top: plot of the safe and disruptive regions of the operational space in JET with the ILW. The colour code represents the posterior probability of the classifier. The black circles are all the non-disruptive shots (10 random time slices for each shot). The red squares are the data of the disruptive shots at the time slice when the predictor triggers the alarm. The blue crosses are the false alarms. Bottom: zoom of the most relevant boundary region.**

### *7.2 Adaptive approach to efficient training*

JET database with the ILW has been used to train the SVM as described in the previous subsection. For continuity with the previous literature, the locked mode and internal inductance signals have been provided as input to the

SVM. Indeed various studies performed in the past have shown the importance of these two quantities in predicting the occurrence of a disruptions not only on JET but also on other devices. The posterior probabilities have then been calculated as indicated in Section 2.2. The adaptive training has been performed for a whole range of threshold probabilities. It turns out that the probability value, which provides the best performance, is 60%. Therefore the model trained with this threshold is the one whose results have been reported in the paper. It is worth mentioning that for an interval of 10% around this 60% value, the models give all almost exactly the same results. So the choice of the threshold is not too critical for the purpose of the present paper, the identification of a manageable formula to describe the boundary between safe and disruptive regions of the operational space. The results of the systematic tests performed are reported in Appendix A.

The curve level plots of the posterior probability obtained are reported in Figure 8. The curve in light blue represents the equation derived with SR via GP (see later). The safe and disruptive regions are well separated in the plane of the locked mode and internal inductance. The clear separation is confirmed by the results in terms of success rate and false alarms reported in Table VI. From Table VI it is easy to appreciate the extremely good performance of the probabilistic SVM.

**Table VI. The results reported in the row Training refer to the ones obtained by the adaptive training. The ones in the row called Test have been obtained by reapplying the final model at the end of the last campaign back to the entire set of data.**

| Model | Succes Rate | Tardy | Early | Missed | False | Missed + Tardy |
|---|---|---|---|---|---|---|
| TRAINING | 96.2 % (180/186) | 2.7 % (5/186) | 0.5 % (1/186) | 0.5 % (1/186) | 3.9 % (40/1016) | 3.2 % (6/186) |
| TEST | 97.9 % (183/187) | 2.1 % (4/187) | 0 % (0/187) | 0 % (0/187) | 2.8 % (29/1020) | 2.1 % (4/187) |

The methodology, described in the Section on Symbolic Regression, has then been applied to the model obtained at the end of the adaptive training. The following model has been retained as a good compromise between complexity and accuracy:

$$y(x) = a_0 \exp(a_1 x^{a_2}) \qquad (10)$$

Where y is the locked mode expressed in $10^{-4}$ Tesla, x the internal inductance and the coefficients assume the values:

$$\begin{aligned} a_0 &= 5.4128 \pm 0.0031; \\ a_1 &= -0.11614 \pm 0.00085; \\ a_2 &= 2.21 \pm 0.011; \end{aligned} \qquad (11)$$

The performance of the previous equation, in terms of the usual figures of merit adopted to qualify predictors, reproduce very well the one of the original model as can be appreciated from Table VII.

**Table VII: The figures of merit obtained using equation (10).**

| Probability Thershold | Success rate | Tardy | Early | Missed | False |
|---|---|---|---|---|---|
| 60 | 97.9 % (183/187) | 2.1 % (4/187) | 0 % (0/187) | 0 % (0/187) | 2.8 % (29/1020) |

Comparing Tables VII with Table VI, it is possible to see how the obtained equation reproduces almost exactly the performance of the original model derived by training the probabilistic SVM. In graphical terms, equation (10) is shown in light blue in Figure 8; from the plots of this figure, it easy to appreciate how the analytical formula obtained with the proposed methodology follows almost exactly the 60 % curve level of the probabilistic SVM. Therefore, reformulating the equation of the boundary, in a more interpretable way than the output of the SVM, does not imply any significant loss of information in this case. In addition to the good performance, it must be appreciated how equation (10) represents a major simplification compared to the sum of Gaussians centred on the support vectors, the model of the original SVM training. From the point of view of the physics interpretation, equation (10) shows how the critical amplitude of the locked mode depends on the internal inductance and therefore on the current profile. In particular, more peaked profile can tolerate a higher level of the locked mode before disrupting. This evidence is not in contrast with the treatment proposed in [30], where it is argued that the amplitude of the locked mode is the important quantity to interpret the boundary between the safe and disruptive regions of the operational space (see next section). In any case, independently from

the details of the physics involved, it is clear from equation (10) and the experimental evidence of Figure 8 that a simple threshold in the locked mode, the criterion traditionally used on JET and other devices to launch alarms, is not a the best choice to maximize the performance of predictors.

## 8 Deployment of the proposed approach in support to model building

In the previous section, an explorative case has been described. Data driven models are derived and tested until the best one is selected. In this section, the same tools are applied to the assessment of the quality of already devised models. For the present example, therefore, guidance, in particular with regard to the signals to be used as inputs to the predictors, is obtained from already proposed empirical models. Then a traditional training of the SVM has been implemented.

A first attempt at deriving a reasonable equation separating the safe and disruptive regions of the operational space has been tried using the variables at the basis of the Hugill and beta limit plots. As described in section 1, these traditional representations have practically zero predictive power for JET with the ILW, at least for the campaigns considered in this study. On the other hand, one could ask whether the poor success rates are the consequence of the simple mathematical form of the equations, even if the variables have a high information content and could be good inputs for a classifier. To falsify this hypothesis, the developed methodology has been systematically applied to the quantities, which appear in both representations. Unfortunately, the results have been very negative. The success rate has never been sufficient and the rate of false alarms might even reach 40 %. Therefore, we have to conclude that the quantities entering the Hugill and beta limit plots do not constitute an effective set of features to perform prediction in JET with the ILW.

The main reason for the poor performance of the Hugill and beta limit representations resides in the fact that they do not include the locked mode and the internal inductance among the inputs. The locked mode and the internal inductance signals are really much more informative quantities for disruption prediction than the ones use in the Hugill and beta limit plots. This is also confirmed by a recent model developed for the level of the mode locked leading to disruptions in various Tokamaks [30]. In this study, the amplitude of the locked mode, considered the consequence of magnetic islands locked to the wall, is studied in JET, ASDEX-U and COMPASS. The simple locking of the magnetic configuration to the wall is not deemed a sufficient condition per se to trigger disruptions; the amplitude is proposed to

be the real quantity of relevance. Based on theoretical considerations involving the Chirikov criterion, the amplitude of the island size required to trigger a disruption was determined. A scaling for the value of the locked mode amplitude, considered the limit for the triggering a disruption, was derived by considering its value at the time of the beginning of the current quench. The resulting equation determining the threshold for the occurrence of a disruption is found to be:

$$B_{ML}(r_c) = c \cdot I_P^{a_I} \cdot a^{a_a} \cdot q_{95}^{a_q} \cdot li(3)^{a_{li}} \cdot \rho_c^{a_\rho} \qquad (12)$$

Where $B_{ML}$ is the amplitude of the locked mode, $c$ is a constant, $a_i$ are regression

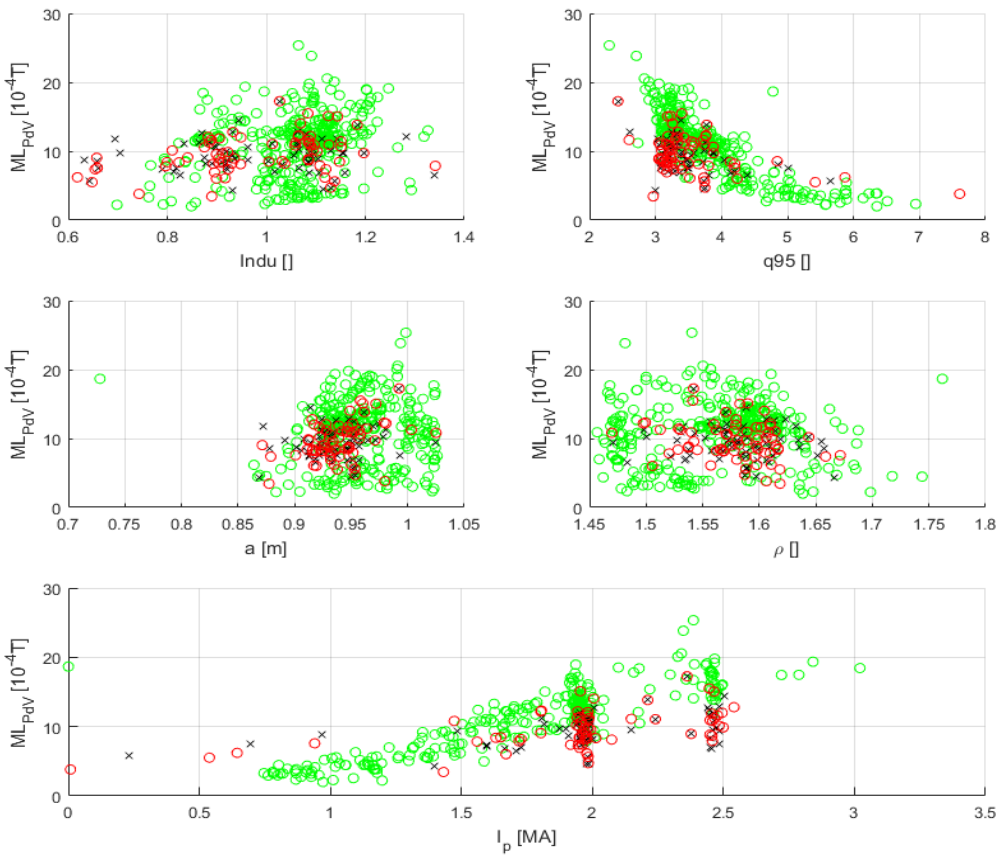


**Figure 9. Plots of the critical value of the locked mode, predicted by equation (12), versus the various quantities used in the regression. Red circles: critical values of the locked mode at moment of the alarm. Black squares: critical values of the locked mode at the time 15 ms before the beginning of the current quench. Green circles: critical values of the locked mode at the flat top of the same discharges. Before the beginning of the current quench the overlap between the values of the locked mode is almost complete**

coefficients, $I_p$ is the plasma current,$q_{95}$ the safety factor at 95% of the radius, $a$ is the minor radius, $l_i$ is the internal inductance and $\rho_c$ the distance between the plasma centre and the location of the magnetic loops measuring the amplitude of the locked mode.

For the database analysed in this paper, equation (12) has been calculated using for the parameters the values suggested in [30] and reported in the following:

$$c = 8.5;\ a_I = 1.07;\ a_a = -1.1;\ a_q = -1.2;\ a_{li} = 1.2;\ a_\rho = -2.8$$

Unfortunately equation (12) does not fit well the data of campaigns C29-C31 of JET with the ILW. This can be appreciated by inspection of the plots reported in Figure 9. For the 170 disruptions considered in the present study, the plots of Figure 9 report the critical value of the locked mode, as predicted by equation (12), versus all the regressors. The values of the critical locked mode has been calculated with equation (12) with the values of the parameters suggested in [30] at the following times: at the time slice of the alarm (when the experimental locked mode reaches the critical value predicted by the equation) and 15 ms before the beginning of the current quench. The estimates of the critical threshold have also been calculated for the flat top, safe phase of the same discharges. From the plots reported, it can be easily seen that before the beginning of the current quench there is full overlap between the estimates for the flat top and the pre-disruptive time slices. Therefore, these estimates are not extremely useful for prediction. This has been verified by testing the performance of the

**Table VIII: The traditional figures of merit to assess the performance of predictors for the case of equation (12)**

| Succes Rate | Tardy | Early | Missed | False | Missed + Tardy | Mean [ms] | Std [ms] |
|---|---|---|---|---|---|---|---|
| 51.85 % (70/135) | 26.67 % (36/135) | 0 % (0/135) | 21.48 % (29/135) | 1.96 % (20/1020) | 48.15 % (65/135) | 184 | 349 |

critical value of the locked mode proposed in [30] for the entire C29-C31 campaigns (for a total of 170 disruptive shots and 987 safe discharges). Again, all the time slices at plasma current higher than 750 kA have been included in the analysis. The final statistics are reported in Table VIII, from which it can be appreciated how the success rate is certainly less than satisfactory.

Even if equation (12) does not seem to provide very useful information for prediction in JET with the ILW, a least for the campaigns analysed, the set of quantities proposed contain signals which are unquestionably very important. In particular the locked mode and the internal inductance have proved to be essential also in the analysis reported in Section 7. Therefore, it can be argued that it is the power law form of equation (12), which is not adequate to model JET data. The probabilistic SVM has therefore been applied to the regressors used in equation (12) to derive the critical value of the locked mode signal. The models derived present very good performance, as can be appreciated by inspection of the table in Appendix B. In this case, using a threshold of 80% in probability seems to provide a very good compromise between success rate and false alarms. For this choice of threshold

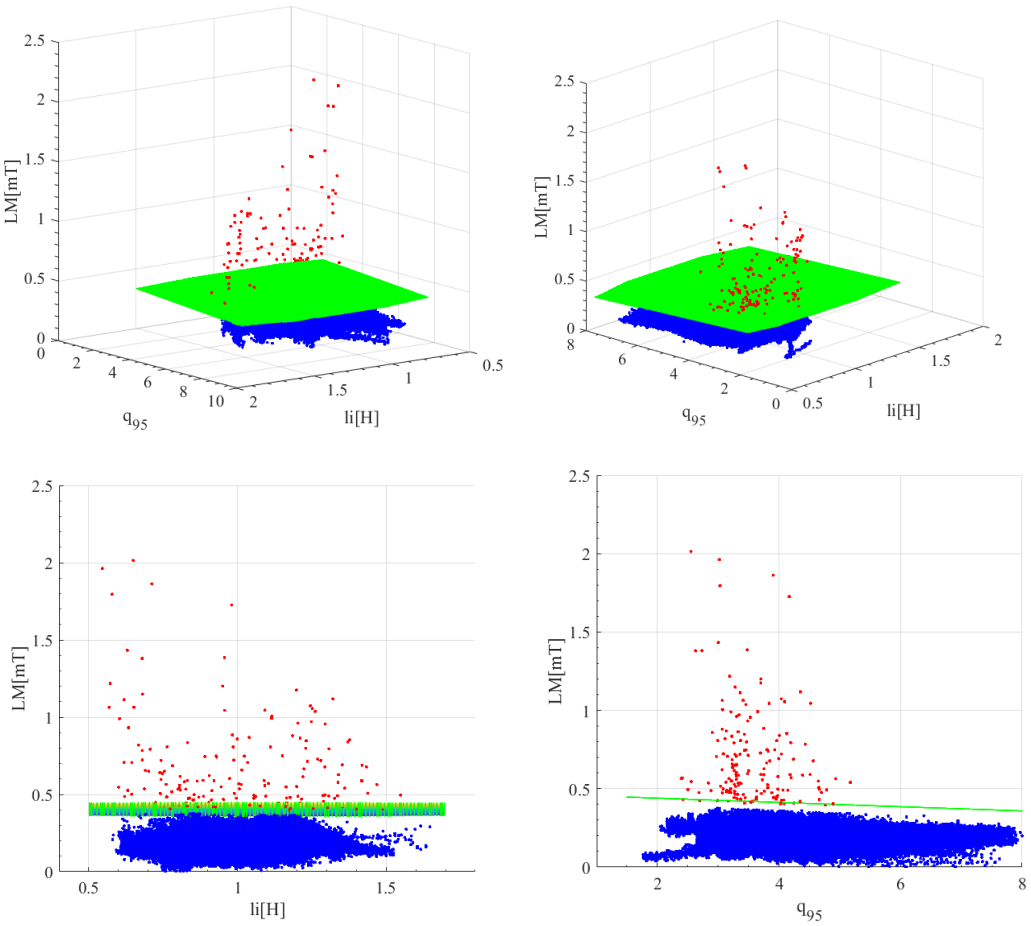


**Figure 10. The boundary between the safe and disruptive regions of the operational space in the three dimensions x, y, z. Only the first disruptive point has been reported to help visualizing the behaviour of data.**
**In green the hypersurface obtained with symbolic regression and non linearly fitted to the data. In red the first disruptive point, in blue all the non disruptive ones.**

probability, the success rate is about 95% and the false alarms are 5%. Therefore the model performs slightly worse than the one of equation (10) but it is very competitive.

The main reason for the better performance of the machine learning models, compared to equation (12), resides as expected in the functional form of the final equations. Unfortunately, it is not possible to apply the technique proposed in this paper also for such a high number of regressors. Indeed the dimensionality of equation (12) is too high and the number of disruptions insufficient (by orders of magnitude). In such a number of dimensions, the points are to sparse in the hyperspace and it is not possible to reconstruct a boundary with any physical sense. On the other hand, the number of really relevant quantities for JET is three: locked mode, internal inductance and $q_{95}$. The others, such as the distance between the plasma boundary and the location of the coils measuring the locked mode, are introduced in [30] to obtain a multimachine scaling, which is not the objective of the present studies. In three dimensions, it is still possible to apply the proposed methodology of SR via GP with 170 disruptions as examples. Deploying again the probabilistic SVM, it has been found that the equation for the model with threshold 80% provides again a good compromise between success rate (94%) and false alarms (5.2%). A graphical representation of the boundary between the safe and disruptive regions of the operational space in three dimensions is shown in Figure 10.

Using the 80 % threshold model, an advanced application of symbolic regression has been deployed to obtain the equation of the boundary between the safe and disruptive regions of the operational space. The original range of the variables (including disruptive and non-disruptive samples) is: ML $\epsilon$ [0.0030, 3.1] mT, $q_{95}$ $\epsilon$ [2.07, 7.81], li $\epsilon$ [0.45, 1.64]. The function to be identified express the critical locked mode amplitude as a function of the internal inductance and $q_{95}$, i.e f(x,y)=LM( li, $q_{95}$). The obtained equation reads:

$$\mathrm{LM} = 0.475 - 0.017 \cdot y^{0.95} - 0.014 \cdot x^{1.00} \quad (13)$$

where

$$x = li[H]$$
$$\mathrm{y} = \mathrm{q95}$$
$$\mathrm{z} = \mathrm{LM[mT]}$$

The plots of Figure 10 show how equation (13) fits the boundary between the safe and disruptive regions of the operational space.

**Table IX: Performance of the p=0.8 probability hypersurface of SVM and of the nonlinearly fitted hypersurface from GP at classifying non disruptive pulses.**

| | Correctly classified | False |
|---|---|---|
| SVM | 964/987=0.98 | 23/987=0.02 |
| GP | 964/987=0.98 | 23/987=0.02 |

To confirm the quality of the obtained results, equations (13) has been deployed to classify the experimental points used to train the original probabilistic SVM. The good quality of the obtained results can be appreciated by inspection of Tables IX and X. It is interesting to note that the equations obtained with Symbolic Regression via Genetic Programming have a success rate that perfectly matches the one of the original probabilistic SVM.

Therefore, the results of the investigation with the proposed combination of machine learning tools indicate that the quantities considered in the model of equation (12) are very informative about disruptions. On the other hand, the equation has a different form and in particular it cannot be represented by a simple power law, as assumed implicitly in [30] by applying log regression to the experimental data. To interpret the results the basics physics explanations will probably have to be revisited. In practical terms. Equation (12) indicates that the configuration stability becomes more delicate as the internal inductance and the $q_{95}$ increase, since at higher values of these quantities the plasma can tolerate a smaller value of the locked mode before disrupting.

**Table X: Performance of the p=0.8 probability hypersurface of SVM and of the nonlinearly fitted hypersurface from GP at classifying disruptive pulses.**

| | Correctly classified | Early | Missed | Tardy |
|---|---|---|---|---|
| SVM | **156/170=0.917** | 0/170=0.0 | 2/170=0.012 | 12/170=0.071 |
| GP | **156/170=0.917** | 0/170=0.0 | 2/170=0.012 | 12/170=0.071 |

## 9 Conclusions

In this paper, it is shown for the first time how it is possible to derive in full generality an equation for the boundary between safe and disruptive regions of the operational space directly from the classification provided by a machine learning tool, namely a probabilistic

SVM. This goal is achieved by an original application of Symbolic regression via Genetic Programming. The performance of the derived equations, in terms of success rate, are practically the same as the original machine learning tools. Therefore, the critical aspect to obtain valid equations is the quality of the statistical basis used to train the machine learning tools.

The data driven models derived with the methodology described in the paper, clearly outperform by a factor, if not by one order of magnitude, traditional empirical models based on representations such as the Hugill or the beta limit plots. The derived relations for the boundary are also orders of magnitude easier to interpret that the typical equations obtained from traditional SVM. Therefore, the developed techniques can be tuned to find the best trade-off between complexity and realism; the derived models are of a manageable complexity and, at the same time, do not oversimply the problem at the expenses of poor success rates like the power laws recently proposed. Moreover, by appropriate selection of the Symbolic Regression basic function, it is possible to obtain equations with physical meaning, which can be compared with theory or used to guide model developments. It is worth emphasizing that the formulation of the equations in more physically meaningful form does not cause any reduction in the success rate of classification. Therefore, these equations can be usefully deployed also in real time networks for the actual prediction of avoidance of disruptions.

It is important to notice that the analysis presented show very clearly how, at least in the features space investigated in the paper, the boundary between the safe and disruptive regions of the operational space can have a quite different mathematical form, depending on the number of regressors used. In the case of two independent variables, the equation of the boundary is an exponential whereas in a three dimensional space becomes a linear equation. This emphasizes the importance of the methodology proposed, which does not force *a priori* the models to have any predefined mathematical form. Moreover, power laws have not proved to be very useful expressions for the boundary on the operational space in JET with the ILW.

**Aknowledgements**

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. This work was also partially funded by the Spanish Ministry of Economy and Competitiveness under the Project No. ENE2015-64914-C3-1-R

**References**

[1] C.R.Hadlock "*Six causes of Collapse*" Mathematical Association of America Washington 2012

[2] R.Wenninger et al "*Power Handling and Plasma Protection Aspects that affect the Design of the DEMO Divertor and First Wall*" submitted for publication in Proceedings of 26th IAEA Fusion Energy Conference

[3] J.Wesson "*Tokamaks*" Oxford University Press 2011

[4] A.Murari et al Nucl. Fusion 57 (2017) 016024 (11pp) doi:10.1088/0029-5515/57/1/016024

[5] A.Murari et al Nuclear Fusion, Volume 49, Number 5 April 2009 doi.org/10.1088/0029-5515/49/5/055028

[6] A.Murari et al Nuclear Fusion, Volume 48, Number 3 February 2008 doi.org/10.1088/0029-5515/48/3/035010

[7] G.Rattà et al Nuclear Fusion, Volume 50, Number 2 January 2010 doi.org/10.1088/0029-5515/50/2/025005

[8] Y.Zhang et al Nuclear Fusion, Volume 51, Number 6 May 2011 doi.org/10.1088/0029-5515/51/6/063039

[9] J. Vega, S. Dormido-Canto, J. M. López, A. Murari, J. M. Ramírez, R. Moreno, M. Ruiz, D. Alves, R. Felton and JET-EFDA Contributors. "Results of the JET real-time disruption predictor in the ITER-like wall campaigns". Fusion Engineering and Design 88 (2013) 1228-1231.

[10] J. Vega, R. Moreno, A. Pereira, S. Dormido-Canto, A. Murari and JET Contributors. "Advanced disruption predictor based on the locked mode signal: application to JET". 1st EPS Conference on Plasma Diagnostics. April 14-17, 2015. Book of abstracts. Frascati, Italy.

[11] .Vega, A. Murari, S. Dormido-Canto, R. Moreno, A. Pereira, G. A. Rattá and JET Contributors. "Disruption Precursor Detection: Combining the Time and Frequency Domains". Proc. of the 26th Symposium on Fusion Engineering (SOFE 2015). May 31st-June 4th, 2015. Austin (TX), USA

[12] J. Vega, R. Moreno, A. Pereira, S. Dormido-Canto, A. Murari and JET Contributors. "Advanced disruption predictor based on the locked mode signal: application to JET". Proceedings of Science . ECPD 2015, 028

[13] B.Cannas at al *Nuclear Fusion 53 093023, 2013* doi.org/10.1088/0029-5515/53/9/093023

[14] B.Cannas et al *Plasma Phys. Control. Fusion 57 125003, 2015* doi.org/10.1088/0741-3335/57/12/125003

[15] A.Murari et al Nuclear Fusion, Volume 53, Number 3 February 2013 doi.org/10.1088/0029-5515/53/3/033006

[16] J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, A. Pereira, A. Acero and JET-EFDA Contributors. "Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks". Nuclear Fusion. 54 (2014) 123001 (17pp).

[17] S. Dormido-Canto, J. Vega, J. M. Ramírez, A. Murari, R. Moreno, J. M. López, A. Pereira and JET-EFDA Contributors. "Development of an efficient real-time disruption predictor from scratch on JET and implications for ITER". Nuclear Fusion. 53 (2013) 113001 (8pp).

[18] 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 Disruption Mitigation, Avoidance and Classification on JET"* submitted to Nuclear Fusion

[19] Steinwart, Ingo; and Christmann, Andreas; *Support Vector Machines*, Springer-Verlag, New York, 2008. ISBN 978-0-387-77241-7

[20] Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola et al. (ed.), Advances in Large Margin Classifiers. MIT Press, Cambridge, MA, 2000.

[21] C. Lin, C. Hsu. "A comparison of methods for multiclass support vector Machines". IEEE Transactions on Neural Networks 13 (2002)

[22] J. Weston, C. Watkins, "Support Vector Machines for multi-class pattern recognition". Proceeding of the Seventh European Symposium on Artificial Neural Networks, 1999

[23] Schmid M. and Lipson H. 2009 *Science* **324** 81–5

[124] Koza J.R. 1992 "*Genetic Programming: on the Programming of Computers by Means of Natural Selection"* (Cambridge:MIT Press)

[25] Burnham K.P. and Anderson D.R. 2002 "*Model Selection and Multi-Model Inference: a Practical Information-Theoretic Approach"* 2nd edition (New York: Springer)

[26] Bates, Douglas, and Watts, Donald, "*Nonlinear Regression Analysis and Its Applications*", Wiley,1988

[27] A. Murari et al 2013 *Nucl. Fusion* **53** 043001 doi:10.1088/0029-5515/53/4/043001

[28] A.Murari et at Nuclear Fusion, Volume 56, Number 2 (2015) doi:10.1088/0029-5515/56/2/026005

[29] A. Murari et al Plasma Physics and Controlled Fusion (2015),**57** (1), doi: 10.1088/0741-3335/57/1/014008

[30] P.C. de Vries et al Nuclear Fusion, Volume 56, Number 2 December 2015 doi.org/10.1088/0029-5515/56/2/026007

## APPENDIX A: Performance of the MAPP for various choice of the triggering window: database of JET with the ILW.

**Table A1. Main figures of merit of MAPP quality using the posterior probability to decide whether to trigger an alarm. This adaptive predicators have been implemented retraining after one time slice detected as disruptive.**

| Soglia post prob DISR | Succes Rate | Missed | False | Early | Tardy | Mean [ms] | Std [ms] |
|---|---|---|---|---|---|---|---|
| 20 | 96.77 | 2.69 | 4.72 | 0.54 | 2.15 | 336 | 345 |
| 30 | 96.77 | 2.69 | 4.72 | 0.54 | 2.15 | 336 | 345 |
| 40 | 96.77 | 2.69 | 4.53 | 0.54 | 2.15 | 335 | 345 |
| 50 | 96.77 | 3.23 | 3.74 | 0.00 | 2.69 | 326 | 334 |
| 60 | 96.77 | 2.69 | 3.84 | 0.54 | 2.15 | 334 | 345 |
| 70 | 96.77 | 3.23 | 3.35 | 0.00 | 2.15 | 330 | 342 |
| 80 | 94.09 | 5.38 | 2.07 | 0.54 | 4.30 | 321 | 344 |

**Table A2. Main figures of merit of MAPP quality using the posterior probability to decide whether to trigger an alarm. This adaptive predicators have been implemented retraining after two consecutive time slices detect a disruption.**

| Soglia post prob DISR | Succes Rate | Missed | False | Early | Tardy | Mean [ms] | Std [ms] |
|---|---|---|---|---|---|---|---|
| 20 | 96.77 | 2.69 | 5.12 | 0.54 | 2.15 | 335 | 345 |
| 30 | 96.77 | 2.69 | 4.53 | 0.00 | 2.15 | 335 | 345 |
| 40 | 96.24 | 3.23 | 3.44 | 0.54 | 2.69 | 331 | 344 |
| 50 | 96.24 | 3.23 | 3.25 | 0.54 | 2.15 | 330 | 341 |
| 60 | 96.24 | 3.76 | 3.65 | 0.00 | 3.23 | 333 | 344 |
| 70 | 94.62 | 4.84 | 2.66 | 0.54 | 3.76 | 324 | 343 |
| 80 | 93.55 | 6.45 | 1.77 | 0.00 | 5.38 | 317 | 342 |

**Table A3. Main figures of merit of MAPP quality using the posterior probability to decide whether to trigger an alarm. This adaptive predicators have been implemented retraining after three consecutive time slices detect a disruption.**

| Soglia post prob DISR | Succes Rate | Missed | False | Early | Tardy | Mean [ms] | Std [ms] |
|---|---|---|---|---|---|---|---|
| 20 | 95.70 | 3.76 | 4.43 | 0.54 | 3.23 | 332 | 344 |
| 30 | 94.09 | 5.91 | 3.25 | 0.00 | 4.84 | 323 | 340 |
| 40 | 94.09 | 5.38 | 2.95 | 0.54 | 4.30 | 325 | 342 |
| 50 | 94.62 | 5.38 | 2.27 | 0.00 | 4.30 | 324 | 340 |
| 60 | 92.47 | 6.99 | 2.07 | 0.54 | 6.45 | 309 | 334 |
| 70 | 92.47 | 7.53 | 1.87 | 0.00 | 6.45 | 319 | 341 |
| 80 | 93.01 | 6.99 | 1.18 | 0.00 | 4.84 | 321 | 342 |

**APPENDIX B: Performance of the MAPP for various threshold percentages using as inputs: ML[mT], $q_{95}$,Li[H],Rmag~ρc,a[m],I[MA].**

**Table B1 Results of the Classification (170 disruptive pulses and 987 non disruptive ones) using as inputs ML[mT], $q_{95}$,Li[H], without standardizing the variables. The disruption probability is indicated by pd.**

| **pd=0.5** | | | | | |
|---|---|---|---|---|---|
| Non Disruptive | Correct | False | Early | Missed | Tardy |
| | 0.963 =950/987 | 0.037 =37/987 | | | |
| Disruptive | | | | | |
| | 0.929 =158/170 | | 0.006 =1/170 | 0.012 =2/170 | 0.053 =9/170 |
| **pd=0.6** | | | | | |
| Non Disruptive | Correct | False | Early | Missed | Tardy |
| | 0.970 =957/987 | 0.030 =30/987 | | | |
| Disruptive | | | | | |
| | 0.924 =157/170 | | 0.006 =2/170 | 0.012 =2/170 | 0.059 =10/170 |
| **pd=0.8** | | | | | |
| **Non Disruptive** | **Correct** | **False** | **Early** | **Missed** | **Tardy** |
| | **0.98 =964/987** | **0.002 =23/987** | | | |
| **Disruptive** | | | | | |
| | **0.917 =156/170** | | **0.0 =0/170** | **0.012 =2/170** | **0.071 =12/170** |