WPJET1-PR(17) 18602

A Murari et al.

A Syncretic Approach to Knowledge Discovery for the Natural Sciences

Preprint of Paper to be submitted for publication in
Data mining and knowledge discovery

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.

A Syncretic Approach to Knowledge Discovery for the Natural Sciences

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

by A.Murari[1], E.Peluso[2], M. Lungaroni[2], P.Gaudio[2] 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) Associazione EURATOM-ENEA - University of Rome "Tor Vergata", Roma, Italy

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

Abstract

Classification, which means discrimination between examples belonging to different classes, is a fundamental aspect of most scientific applications. Machine Learning (ML) tools have proved to be very performing in this task, in the sense that they can achieve very high success rates. On the other hand, the "realism" and interpretability of their models are very low, resulting often in modest increases of knowledge and limited applicability. In this paper, a methodology is described, which, by applying ML tools directly to the data, allows formulating new scientific models that describe the actual "physics" determining the boundary between the classes. The proposed technique consists of a stacked approach of different ML tools, each one applied to a specific subtask of the scientific analysis; all together they combine all the major strands of machine learning, from rule based classifiers and Bayesian statistics to genetic programming and symbolic manipulation. To take into account the error bars of the measurements, an essential aspect of any scientific form of inference, the novel concept of the Geodesic Distance on Gaussian manifolds is adopted. The characteristics of the methodology have been investigated with a series of systematic numerical tests, for different types of classification problems. The potential of the approach to handle real data has been tested with various experimental databases. The obtained results indicate that the proposed method permits to find a good trade-off between accuracy of the classification and complexity of the derived mathematical equations. Moreover, the derived models can be tuned to reflect the actual phenomena, providing a very useful tool to bridge the gap between data, machine learning tools and scientific theories.

1 Knowledge Discovery in the natural sciences with particular attention to Big Physics experiments

Nowadays the complexity of the problems, investigated in many fields of science, is such that it can become difficult, if not impossible, to describe the phenomena to be studied with theoretical models based on first principles. A typical example in physics is the case of magnetic confinement thermonuclear fusion, whose plasmas are so complex that various levels of modelling (particle, fluid, kinetic etc) coexist without providing a satisfactory description of many aspects of the physics [1]. On the other hand, in the last decades much more data has become available, due to the diffusion of new sensors and cheap but powerful computing units. For example, the Big Physics European experiments are affected by a data deluge. At CERN, the ATLAS detector can produce Petabytes of data per year. In its prime the Hubble space telescope managed to send to earth Gigabytes of data per day and the data warehouse of the Joint European Torus exceeds 350 Terabytes. Therefore, the inadequacies of theoretical models and the vast amounts of information available have motivated the development of data driven tools, to complement hypothesis driven theories. In this perspective, various machine learning methods have been developed and to a certain extent applied to the natural sciences. They range from Neural Networks and Support Vector Machines to Fuzzy Logic classifiers; a series of examples from the field of thermonuclear fusion can be found in [2,3,4]. Manifold learning tools, such as Self Organising Maps and Generative Topographic Maps, have provided very good results also in terms of describing the space in which the relevant physics takes place [5,6,7].

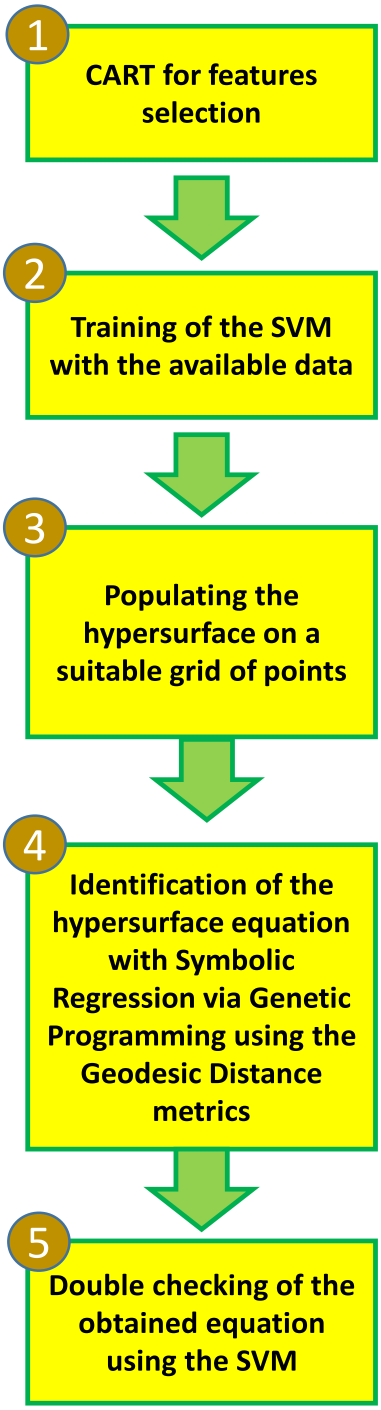
1
CART for features selection
2
Training of the SVM with the available data
3
Populating the hypersurface on a suitable grid of points
4
Identification of the hypersurface equation with Symbolic Regression via Genetic Programming using the Geodesic Distance metrics
5
Double checking of the obtained equation using the SVM

Figure 1: The five main steps of the proposed methodology to express a model identified by SVM in more traditional mathematical notation.

On the other hand, to be really useful, the knowledge discovery process in the natural sciences has to satisfy specific criteria and requirements, which are not necessarily crucial in other applications. In particular, at least four properties are not only desirable but probably essential to the scientific process: a) accuracy of the results b) interpretability of the obtained models c) close relation between the derived equations and the "physics" reality of the phenomena under investigation d) proper treatment of the uncertainties and quantification of the confidence intervals. Even if the traditional data driven tools are providing quite impressive performance in terms of accuracy, their main problem is the mathematical formulation of their models. They have shown the potential to learn very efficiently from the provided examples but their results are expressed in such a way that does not reflect the physics reality behind the phenomena under study. This aspect is quite worrying and has hampered the penetration of many machine learning tools in scientific disciplines such as physics. In conclusions the main problems of traditional machine learning tools in the perspective of applications in the natural sciences are: a) poor "physics fidelity" i.e. excessive discrepancy between the mathematical form of the models and the physical reality of the phenomena investigated b) insufficient estimates of the uncertainties c) difficulties to interpret the results in terms of traditional mathematical formulations d) consequent impossibility to compare the obtained results with mathematical theories based on first principles e) lack of extrapolability of the results.

In order to overcome these limitations, 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 better reflect the reality of the phenomena under study. The techniques, developed in the framework of the activities presented in this paper, address the basic goal of classification. This is a very important task in many scientific applications, both "per se" and as a preliminary step to subsequent investigations. The objective of the analysis, in scientific applications of classification, consists of deriving a mathematical formula for the boundary between the classes, describing the actual physics or chemistry behind the problem. The main idea informing this work resides therefore in combining the learning capabilities of the machine learning tools with the "fidelity" and interpretability of more traditional mathematical formulations, for a more realistic description of the boundaries between classes.

The proposed methodology covers the entire knowledge discovery process, from the feature extraction to the final assessment of the quality of the derived models. A flow chart of the main steps of the proposed technique is provided in Figure 1.

The feature extraction phase is performed with a new evolution of Classification and Regression Trees (CART), the so called noise–based ensemble. The CART approach is particularly useful in this subtask due to the limited computational burden and the high level of interpretability of the results. It is worth pointing out that the issue of the noise and the errors in the measurements is taken into account starting already at this stage, by the original method of the noise-based ensemble, as illustrated in Section 2.

The actual classification step is then based on Support Vector Machines (SVM), whose mathematical background is summarised in the Section 3, including a probabilistic version very important to quantify the confidence in the results. The choice of SVM is mainly due to their structural stability, their capability to maximize the safety margins in the classification. Given the high accuracy of SVM, the equation of their hypersurface in the original space can be considered an excellent approximation of the boundary between the classes. On the other hand, their mathematical representation of the boundary is extremely non intuitive (see Section 3). Indeed referring to complex systems of the complexity investigated in modern day natural sciences, the equations of the hypersurface can easily comprise hundreds of support vectors and therefore the equation of the hypersurface contains an equal number of addends. More importantly, in addition to presenting serious problems for human understanding, the formulation of the boundary equation has typically no relation with the actual dynamics of the phenomena under study. It has indeed been shown, with many numerical examples (see Section 7 and Appendix A), that the models provided by SVM bear absolutely no resemblance to the ones generating the data. A simple methodology has already been proposed and applied to complex problems, to recover the equation of the boundary in the case of linear kernels [8]. In this paper, a new technique is developed, which is fully general. Indeed the proposed method can be applied to SVM with any type of kernel and even to probabilistic versions; therefore it has a much wider range of applications than the more traditional techniques. This aspect is very important in many scientific fields, whose phenomena cannot be simply modelled by linear tools or logistic regression.

To formulate the outputs of SVM in a way suitable for scientific investigations, extensive used is made of Symbolic Regression (SR) via Genetic Programming (GP); these tools are therefore described in Section 4. Symbolic regression is basically used to fit points on the hypersurface found by the SVM, which is considered the boundary between the classes. More formally, in this step the analysis task to be performed consists of (see also Section 3):

given a decision surface $D(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i y_i H(\mathbf{x}_i, \mathbf{x})$ *=0 found by the SVM;*

find an approximate equation $f(x_1,....x_n) = 0$ *which is expressed in a mathematical form suitable to describe the actual physics of the phenomenon.*

To take into account the error bars of the measurements, the formalism of the Geodesic Distance on Gaussian manifolds (GD) has been adopted. Basically this has been inserted in the symbolic regression step: the fitness function in this phase of the method is calculated using the GD. The implementation of the Geodesic Distance on Gaussian manifolds is described in Section 5.

The actual combination of the various tools to provide the equation of the boundary between two classes in a physically relevant form is described in detail in Section 6. The results of a systematic series of numerical tests, proving the potential of the proposed methodology, are the subject of Section 7 and AppendixA. Some examples of application to experimental databases, covering completely different scientific disciplines, are provided in Section 8. Discussions and lines of future developments are the subject of the last Section 9.

Before embarking on the technical description of the develop methodology, a few clarification remarks are appropriate. The approach proposed in this paper is aimed at reconciling 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 scientific theories and models. It is therefore worth emphasizing that the objective of the present work is not simply improving interpretability of machine learning tools, on which significant work has already been done [9,10]. The most important aspect indeed is “physics fidelity” i.e. the formulation of the results in mathematical terms which can be compared with basic theories and models of the various scientific disciplines. Therefore, the proposed method must have the potential to derive mathematical expressions, which reflect the underlining dynamics of the phenomena investigated. This means that the approach must be y flexible enough to allow the outputs of machine learning tools to be expressed in sufficiently complex mathematical forms to describe properly the problems to be studied. On the other hand, it must be possible to control overfitting and convergence on models of the appropriate level of complexity. The other essential aspect of the proposed methodology, for relevant investigations in the natural sciences, is the principled treatment of the measurement errors, to obtain reliable confidence intervals in the results. This has been achieved with the development of the concepts of

Information Technology and in particular the Geodesic Distance on probabilistic manifolds. The other important point to notice is that, as can be seen in Figure 1, the proposed methodology involves practically all the major fields of machine learning, from rule-based classifiers to Bayesian statistic, genetic programming and symbolic manipulation. Each technique is deployed to solve a specific aspect of the data driven theory process, to which it is particularly suited. This stacked, syncretic approach to knowledge discovery seems to be particularly promising for applications in the natural sciences, in which it is already finding increasing acceptance (see Section8). It is also worth mentioning that, in the framework of the present study, it has been possible to devise an adaptive from of training, which is very relevant for many scientific applications analysing time series. Indeed in the vast majority of applications in the natural sciences dealing with data streams, the assumption that the data are i.i.d. (independent and identically distributed) is far from being satisfied. Developing adaptive training schemes from scratch is therefore particularly relevant (See Section 8 and Appendix C).

2 Noise-based ensembles of CART classifiers for feature selection

Among the rule-based machine learning tools, the so called Classification and Regression Trees (CART) are the most developed and widespread. They have been widely implemented for constructing prediction models from data [11,12]. Such models are derived directly from the available databases by recursively partitioning the feature space and fitting a simple prediction rule at each partition. The final partitioning, once properly optimised, consists therefore of a series of rules that can be represented graphically by a decision tree. The performance of classification trees are typically quantified in terms of misclassification costs. The algorithms of this family exhaustively search the whole database to determine, for which variable, which value minimizes the total impurity of its the child nodes. To quantify the purity of a node, the version of CART implemented in this paper uses a generalization of the binomial variance called the Gini index. As a metric to split the nodes, the Gini impurity calculates how often a randomly chosen element from the training set would be incorrectly labelled, under the assumption that the labels are allocated as the distribution of labels in the subset. The Gini factor is typically computed by summing the probability p_i of the item being correctly classified by the probability $(1-p_i)$ of the item being wrongly classified:

$$\mathrm{GINI} = \Sigma\, p_i(1-p_i) \qquad (1)$$

Where the sum is extended over the number of classes. The GINI impurity reaches its minimum (zero) when all cases in the node fall into a single target category.

Decision trees are very practical and easy to interpret but present a significant drawback, consequence of their greedy search strategy: their sensitivity to the specific data used for their training. Indeed a small change in the inputs (for example even using a subset of the training data) can imply a major variation in the resulting decision tree and in turn quite different predictions. To overcome this issue and to increase the success rate of the results, it is typically very advantageous to adopt the approach of ensemble rule-based classifiers, based on the concept of weak learners. A 'weak' learner (either classifier or predictor) is just any machine learning tool, which generates models that perform relatively poorly but are computationally simple [11]. The relatively limited computational resources required allow training various versions of such weak learners which can then be pooled (via Bagging, Random Forests etc) together to create a "strong" ensemble classifier. The traditional versions of pooling present two main drawbacks. First, they have to reduce the number of training examples by subsampling. Second, and partly a consequence of the previous issue, they need to train a very high number of weak learners, which increase the computational complexity and reduce the interpretability of their results. These issues have motivated the development of specific methods to build ensembles for applications in the natural sciences.

One of the main issues of the measurements in the experimental sciences is the often high levels of noise. This noise is very difficult to reduce; the sources of noise are many and independent. Even if these uncertainties are a potential issue, they suggest an alternative approach to the method of building ensembles of weak classifiers, which is an innovation proposed in this paper. The idea consists again of collecting ensembles but not with subsets of the original data; on the contrary the various training sets are obtained by the original one summing random noise to the measurements. The random noise is generated from Gaussian distributions with variance equal to the error bar of the measurements. To each realization of the noise corresponds a different weak learner. The number of trees can be increased until the accuracy begins to saturate instead of improving. This approach called Noised-based Ensemble can be applied directly to CART trees. The main advantage is twofold. First, the noise-based ensemble improves significantly the success rate of the classification at a modicum of additional computational expenses. Secondly, the number of trees in the ensemble can be typically kept to a minimum, with a significant advantage in terms of interpretability. Indeed the ensemble can therefore be explored and a typical tree identified, allowing to analyse directly the rules and derive the required information.

In the application described in this paper, the Noise-based Ensembles of CART trees are used for feature selection. They are trained to classify and then a simple inspection of the trees and the rules allow identifying the most significant features, to be used in the following steps of the methodology. An example of this approach potential in real time applications is provided in Section 8.

3 Traditional and probabilistic SVM

In intuitive terms, given a set of input examples, which belong to two different classes, SVM map the inputs into a high-dimensional space through some suitable non-linear mapping [13]. 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 closest points, the *Support Vectors* (SV), of each class. This is achieved by a careful selection of the constraints of a suitable functional to minimize. In the case of non separable problems, the points to classify are projected into a higher dimensional space with the help of suitable kernels. The minimization of the error risk and the maximization of the margins is then performed in this projected space.

SVM therefore basically consist of suitable kernels, which map the inputs into higher dimensional spaces, where the classification becomes a linearly separable problem and can be solved with traditional quadratic programming methods based on Lagrange multipliers.

In mathematical terms, given a training set of ℓ samples $(\mathbf{x}_1, y_1),...,(\mathbf{x}_\ell, y_\ell)$, $x_i \in \Re^n$, for a binary classification problem (*i.e.* $y_i \in \{+1,-1\}$), SVM estimates the following decision function:

$$D(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i y_i H(\mathbf{x}_i, \mathbf{x}) \qquad (2)$$

where $H(\mathbf{x}_i, \mathbf{x})$ is a kernel function and the parameters $\alpha_i, i = 1,...,\ell$ are the solutions of the following quadratic optimization with linear constraints:

maximization of the functional

$$Q(\alpha)=\sum_{i=1}^{\ell}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j H\left(\mathbf{x}_i,\mathbf{x}_j\right) \qquad (3)$$

subject to the constraints

$$\sum_{i=1}^{\ell} y_i\alpha_i=0, \quad 0\le\alpha_i\le\frac{C}{\ell}, i=1,\dots,\ell \qquad (4)$$

where C is a regularization parameter [13].

The data points $\mathbf{x}_i$ associated with nonzero values of the coefficients α_i are called support vectors, which give the name to the technique. Once the support vectors have been determined, the SVM boundary between the two classes can be expressed in the form

$$D(\mathbf{x})=\sum_{\text{support vectors}}\alpha_i y_i H\left(\mathbf{x}_i,\mathbf{x}\right) \qquad (5)$$

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

The rule to classify a feature vector $\mathbf{u}$ as class C_1 or class C_2 is given by:

if sgn (D(**u**)) ≥ 0
u ε C_1
otherwise
u ε C_2
where sgn(t) is the sign function.

At this point a clarification of the terminology is probably appropriate. SVM find a separating hyperplane in the transformed space. On the other hand, the hyperplane is expressed in terms of Support Vectors in the original space, in which the boundary is a hypersurface. Since typically in the natural sciences researchers are interested in equations in the original space, and not in the transformed one, the boundary between the two classes will be indicated with the term hypersurface and not hyperplane in the following. Indeed, another advantage of the SVM is that their results are expressed in terms of the inputs in the original space. The second and third real life examples described in Section 8 adopt this approach of the traditional SVM. On the other hand, the availability of classifiers, which can output a probability, would be extremely useful in most applications. Unfortunately, traditional SVM, as just described, provide only a distance to a hyperplane, in the form reported in equation

(5). Their basic version has therefore to be extended to associate a probability to the outputs of their classification [14,15].One possible solution consists of reformulating the SVM output in terms of a probability with the Bayes rule according to the formula:

$$P(y = 1|D) = \frac{p(D|y=1)P(y=1)}{\sum_{i=-1,1} p(D|y=i)P(y=i)} \qquad (6)$$

In equation (6) D are the data and y indicates the label of one of the classes. P(y=1) is the prior probability and p(D|y=1) is the likelihood. Therefore, to convert the outputs of traditional SVM to probabilities, two quantities have to be determined: the prior probability and the likelihood. In many applications, the natural choice of the prior probability is the percentage of examples seen, up to a certain point in time in the experiments or observations, for the class to which the SVM labels the new example. The most challenging aspect of relation (6) resides in the evaluation of the likelihood. If a solid and reliable estimate of the likelihood is not viable for any reason, theoretical investigations and practical considerations have shown that one advantageous alternative consists of remapping the distance to the hyperplane to a probability by using a sigmoid function [14,15]:

$$P(y = 1|d) = \frac{1}{1+exp(Ad+B)} \qquad (7)$$

In equation (7) A and B are two fitting parameters, whereas *d* is the distance of the examples to the SVM hyperplane. Equation (7) therefore allows converting 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 (7). 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. To obtain the points to be fitted with symbolic regression (see next Section 4), it is sufficient to select the most appropriate probability threshold (typically the one with better performance in terms of success rate). The points at that level of probability are the inputs to the fitting part of the procedure. This solution of fitting a sigmoid is the one used in the first real life example described in Section 8 of the paper.

4 Symbolic Regression via Genetic Programming for physics fidelity

As mentioned in the first Section, this paper describes a technique to present the results of machine learning tools in a mathematical form describing realistically the actual phenomena to be studied. In the case of classification with SVM, this task consists of representing the hypersurface separating the classes in a more realistic way than the sum of hundreds of terms as in (5) or a series of points at the same probability (in the case of probabilistic SVM). To this end, the main tool used is Symbolic regression via Genetic Programming. The methods developed on the one hand allow identifying the most appropriate mathematical expression for the hypersurface without “a priori” hypotheses. In this way therefore the potential of SVM is fully exploited and no unnecessary restrictions are imposed on the form of the solutions. On the other hand, the complexity of the obtained solutions can be controlled, allowing to find the best trade-off between complexity, success rate of classification and realism of the final models, depending on the objectives of the study.

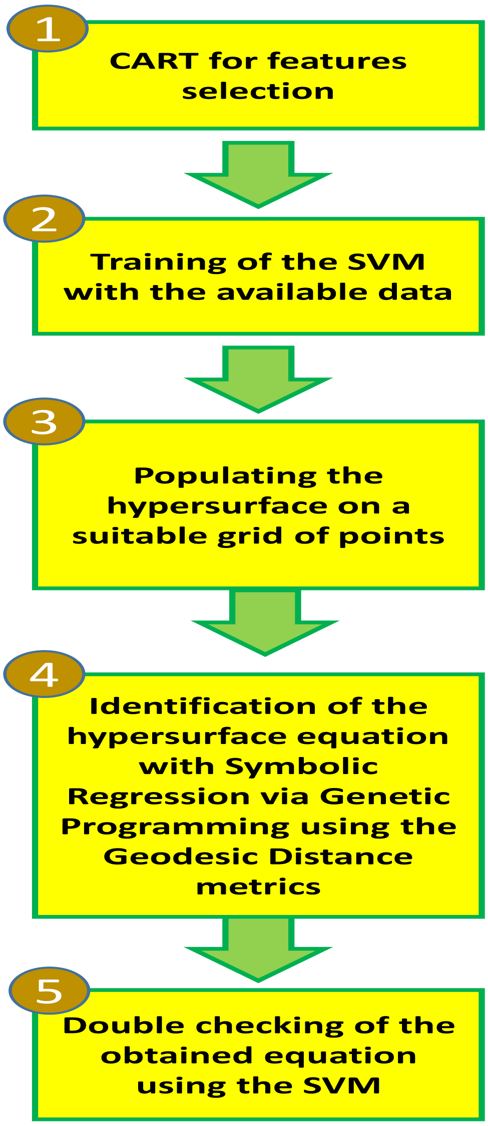
1
CART for features selection
2
Training of the SVM with the available data
3
Populating the hypersurface on a suitable grid of points
4
Identification of the hypersurface equation with Symbolic Regression via Genetic Programming using the Geodesic Distance metrics
5
Double checking of the obtained equation using the SVM

Figure2: The main steps of the proposed methodology to identify the best models without assumptions on their mathematical form.

The method of SR via GP consists of testing various mathematical expressions to fit a given database. The main steps to perform such a task are reported in Figure 2. First of all, the various candidate formulas are expressed as trees, composed of functions and terminal nodes. A simple example of this form of knowledge representation is provided in Figure 3. The function nodes can be standard arithmetic operations and/or any mathematical functions, squashing terms as well as user-defined operators [16,17]. The function nodes included in the analysis performed in this paper are reported in Table I. This representation permits to steer the models towards physics fidelity by proper selecting the basis functions and/or the structure of the trees. Moreover expressing the formulas as trees allows an easy implementation of the next step, symbolic regression with Genetic Programming (GP). Genetic Programs are computational methods able to solve complex optimization problems [16,17]. 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 boundary equation in our case. An appropriate fitness function (FF) is selected to measure how good an individual is with respect to the database. A higher probability to have descendants is assigned to those individuals with better FF. Therefore, the better the adaptation (the value of the FF) of an individual to a problem, the higher is the probability that its genes are passed to its descendants.

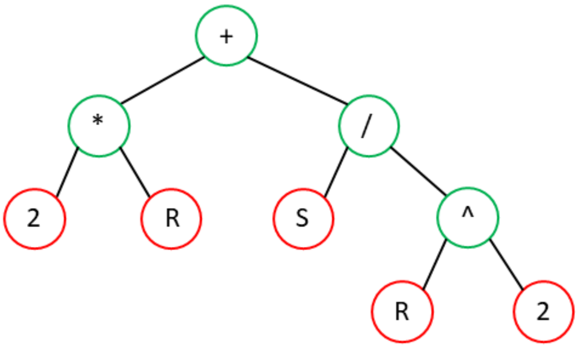

Figure 3: An example of syntax tree structure for the function 2R + (S/R^2). The function operator nodes (green) and the variable or constant nodes (red) are reported.

In more detail, the first step of the method is the generation of the initial population of formulas for the boundary between two classes; then the algorithm assess the quality of each element of the population by evaluating its performance with the metric expressed by the FF. In the following step, 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. This means that better individuals are more likely to have more children than inferior individuals. When a stable and acceptable solution, in terms of 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.

The fitness function is a 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 [18] for the FF. The classic form of the AIC indicator is:

TableI: 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 class	***List***
Arithmetic	*c (constants),+,-,*,/*
Exponential	*exp(x_i),log(x_i),power(x_i, x_j), power(x_i,c)*
Squashing	*logistic(x_i),step(x_i),sign(x_i),gauss(x_i),tanh(x_i), erf(x_i),erfc(x_i)*
Trigonometric	*sine, cosine, hyperbolic sine, hyperbolic cosine*

$$AIC = 2k + n \cdot \ln(RMSE/n) \qquad (8)$$

In equation (8), RMSE is the Root Mean Square Error between the data and the model predictions, k is the number of nodes used for the model and n the number of examples provided, i.e. 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 AIC, as the other indicators used in this work, are to be minimised: the lower the FF the better the model.

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 [18]:

$$BIC = n \cdot \ln\left(\sigma^2_{(\epsilon)}\right) + k \cdot \ln(n) \qquad (9)$$

where$\epsilon = y_{\mathrm{data}} - y_{\mathrm{model}}$ 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. A more sophisticated form of both the AIC and BIC indicators, to take into account the error bars of the measurements using the formalism of the GD, is introduced in the next section. It is worth mentioning that the two indicators, as expressed by equation (8) and (9) or in their more sophisticated version, can be used interchangeably; one in the fitness function and one to check the final results.

The aim of the KLD is to quantify the difference between the computed probability distribution functions, in other words to quantify the information lost when $p(\vec{y}_{model}(\vec{x}))$ is used to approximate $q(\vec{y}_{data}(\vec{x}))$ [18]. The KLD is defined as:

$$KLD(P\|Q) = \int p(x) \cdot ln\big(p(x)/q(x)\big)dx \quad (10)$$

Where the symbols are 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. In our application *p* is the pdf of the data, considered the reference, and *q* the pdf of the model estimates. 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. A detailed overview of SR via GP for scientific applications is provided in [19].

5 Geodesic distance on Gaussian manifolds to include the effects of the error bars

In this section the geodesic distance on probabilistic manifolds is introduced in subsection 5.1. The use of the geodesic distance in the SR is then detailed in subsection 5.2.

5.1 Geodesic Distance

As seen in the previous section, the goal of SR via GP is to extract the most appropriate formulas to describe the available data. To achieve this, typically a quantity somehow proportional to the sum-of-squares of the distances between the data and the model predictions is used in the FF (the RMSE in equation 8 and the variance in equation 9). In this way, SR is implicitly adopting the Euclidean distance to calculate the (dis)similarity between data points and predictions. However the Euclidean distance has a precise geometrical meaning but implicitly requires considering all data as single infinitely precise values. This assumption can be appropriate in other applications but it is obviously not the case in the natural sciences, since all the measurements typically present an error bar. An alternative idea is to use a new distance between data, which would take into account the measurement uncertainties. The additional information provided by this distance renders the final results more robust.

The idea, behind the approach proposed in this paper, consists of considering the measurements not as points, but as Gaussian distributions. This is a valid assumption in many scientific applications, because the measurements are affected by a wide range of noise sources, which from a statistical point of view can be considered random variables. Since the various noises are also typically independent and additive, they can be expected to lead to measurements with a global Gaussian distribution around the most probable value, the actual value of the measured quantity. Each measurement can therefore be modelled as a probability density function (pdf) of the Gaussian type, determined by its mean μ and its standard deviation σ:

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad (11)$$

Modelling measurements not as punctual values, but as Gaussian distributions, requires defining a distance between Gaussians. The most appropriate definition of distance between Gaussian distributions is the geodesic distance (GD), on the probabilistic manifold containing the data, which can be calculated using the Fischer-Rao metric [20]. For two univariate Gaussian distributions $p_1(x|\mu_1,\sigma_1)$ and $p_2\ (x|\mu_2\ ,\sigma_2\ )$, parameterised by their means μ_i and standard deviations $\sigma_i(i\ =\ 1,2)$, the geodesic distance GD is given by:

$$\mathrm{GD}(\mathrm{p}_1||\mathrm{p}_2) = \sqrt{2}\ln\frac{1+\delta}{1\text{-}\delta} = 2\sqrt{2}\tanh^{-1}\delta,\ \text{where } \delta = \left[\frac{(\mu_1\text{-}\mu_2)^2+2(\sigma_1\text{-}\sigma_2)^2}{(\mu_1\text{-}\mu_2)^2+2(\sigma_1+\sigma_2)^2}\right]^{\frac{1}{2}} \qquad (12)$$

The meaning of GD can be appreciated by inspecting Figure 4, which reports the distance between two couples of Gaussian distributions. The distance between the means of the members of the two couples is the same. On the other hand, the Gaussian pdfs of one couple have a standard deviation an order of magnitude higher the other. The distance between the pdfs with higher standard deviation is therefore significantly lower than the one of the more concentrated pdfs, which is intuitively and conceptually correct since they overlap much more. This property of the GD increases the robustness of the results and reduces the risk of overfitting, as verified with a series of numerical tests (see also next subsection).

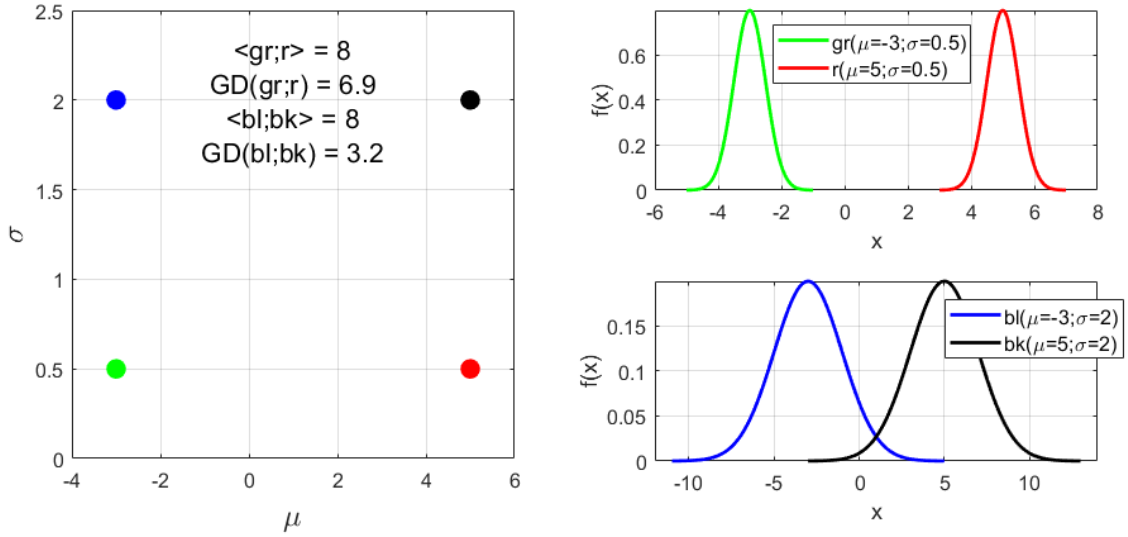

Figure 4. Examples to illustrate how the GD determines the distance between two Gaussians. The two couples of pdf in the figure have the same mean but different σ. The geodesic distance between the two with higher σ is much smaller. GD indicates the geodesic distance and <> the Euclidean distance.

5.2 Use of Geodesic Distance in Symbolic regression

To take into account the measurement errors in a statistically sound way the last step required consists of inserting the GD into the SR. To this end, a good solution has been obtained by replacing the RMSE and variance with the GD in the AIC and BIC criteria, according to the following formulas:

$$AIC = \sum_i GD_i + 2k \qquad (13)$$

$$BIC = n \ln(\sum_i GD_i) + k \ln(n) \qquad (14)$$

where the symbols have the same meaning as in formulas (8) and (9) and the index *i* runs over the entries of the database. It is worth pointing out that this idea of inserting the GD in the FF of the SR is another original development presented in this paper.

Since in the genetic programme, implementing symbolic regression, the GD is to be calculated as the distance between the experimental values and the estimates of the model, the Gaussian parameters μ and σ must be properly chosen. The typical assumption is to take the measured value as the μ, assuming that the average value is the most likely measurement.

For the standard deviation, a reasonable assumption is to adopt the value of the error bars in the experimental points.

With regard to the use of the GD in the AIC and BIC criteria, it should be emphasized that the GD is a statistically sound way to calculate the minimum distance between Gaussian pdfs. In practice, it has been tested on hundreds of thousands of models that the SR using GD is practically never outperformed by the SR using the RMSE or the variance. Moreover, the GD is more robust against outliers. It is indeed a well know statistical fact that the RMSE and variance are not a very robust indicators and are particularly vulnerable to outliers. As an example of these tests, the following equations have been used to generate synthetic data:

$$f_1 = \cos(x_1 \cdot x_2) + \sin x_1^{0.5}$$

$$f_2 = \cos\left(\frac{x_1}{x_2}\right) + 2x_3 \cdot \{1/[1 + exp(-0.8 \cdot x_2)]\}$$

$$f_4 = x_1^{0.8} + \frac{\{1/[1 + exp(-0.6 \cdot x_2)]\}}{x_3}$$

$$f_5 = x_1 \cdot x_2 \cdot exp(-x_3) + 2x_3$$

The range of variations of the independent variables, for the examples reported in the following, is:

$$x_1 = 0.015 \ldots 3.9$$

$$x_2 = 0.044 \ldots 1.97$$

$$x_3 = 0.268 \ldots 2.178$$

Two different types of noise have been implemented: Gaussian noise of zero mean and standard deviation equal to a fixed percentage of the mean value of the functions and a noise with outliers. The distribution of the outliers has been modelled with a second Gaussian with a mean different from zero. The weight of this second Gaussian modelling the outliers can be selected. In general, for various databases and a number of outliers ranging up to 50% of the entries, SR with the GD outperforms systematically the version using the RMSE. SR with GD manages always to approximate the generating functions not worse than the version with RMSE and it provides better results in about 50%of the cases.

6 Combining SVM and Symbolic Regression for Boundary Equations

This Section describes in detail the combination of SVM technology with SR via GP to obtain the equations of the boundary between classes in a form appropriate for scientific investigations. Subsection 5.1 introduces the proposed way to find points on the hypersurface identified by the SVM. Subsection 5.2 describes the use of symbolic regression for the derivation of the actual formula of the boundary between the classes.

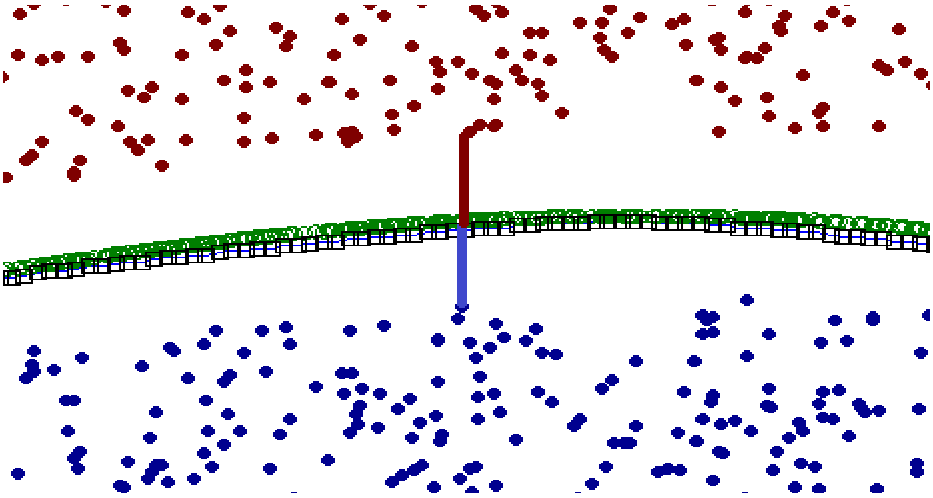

Figure 5: SVM hypersurface points for synthetic, linearly separable data set. For illustrative purposes the distance between the points and the hypersurface has been exaggerated. The blue squares and the green circles represent the points identified as belonging to the hypersurface separating the two classes.

6.1How to find points on the SVM hypersurface

In order to interpret the results produced by the SVM, 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. In the case of probabilistic SVM, obtaining the points on the boundary is technically very simple. The main decision to be taken is the choice of the most appropriate value for the probability threshold to separate the classes; this can be achieved on the basis of the success rate and the objectives of the classification. Obtaining the hypersurface points in the case of a traditional SVM is a bit more involved and requires a specific procedure described in the next part of this subsection. A mesh is built first, with resolution equal or better than the error bars of the measurements used as inputs to the SVM. The limits of the domain is defined by the ranges of variables; therefore, if the problem presents *n* dimensions and *m* grid points are generated for each dimension, the grid will consist of m^n 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 limitations. On the other hand, there are at least two criteria for selecting the number of intervals for different directions more efficiently. The first one consists of allocating more intervals along the direction of stronger curvature. The second, and more useful strategy, consists of allocating a higher number of intervals in the direction of the dependent variable, to make sure that the points selected for the hyper-surface are close enough to the real hyper-plane.

After building the grid, the algorithm starts selecting the 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 considered points on the hypersurface. This assumption is more than reasonable because, 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 in the natural sciences, 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 the line connecting the two sides of the hypersurface is drawn.

It is good practice to repeat the process also starting from the other side of the hypersurface, in order to avoid possible bias in the selection of the points on the hypersurface. An adequate number of points is typically a multiple of the support vectors. One order of magnitudes more points than SVs is a safe choice; attention can also be usefully paid to the fact that the density of the points reflects the density of the SVs in the feature space. In any case, it is easy to increase the number of points up to the number necessary. The main limitation here is mainly computational time not any principle difficulty.

6.2 Deriving the equation of the hypersurface with symbolic regression

Once it has been verified that sufficient 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 described in the previous subsection are on the boundary between the two classes. Therefore the equation of that surface is the equation of the boundary between the two classes.

An efficient way of retrieving the equation of the hypersurface from the points consists of regressing them with SR, using the quantity with a largest dynamic range as the independent variable. The quality of the obtained equation can be assessed first with the statistical indicators described in Section 4. Moreover, an additional and more conclusive test can be performed, exploiting again the trained SVM. In this case, it is indeed possible to generate a series of points from the candidate formula and insert them in the SVM. If the distance from these points and the hypersurface is sufficiently close to zero, it can be confirmed that indeed the equation is a good representation of the boundary between the two classes. As a criterion of closeness to the boundary, typically the value of the error bars of the measurements can be taken: if the points generated by the equation are at a distance from the hypersurface smaller than the error bars, for all practical purposes the obtained equation can be considered a sufficient approximation of the boundary between the two classes.

To take into account the error bars of the measurements, symbolic regression is run with the FF including the geodesic distance, according to equations (13) and (14).

7 Numerical tests of SR via GP for boundary equations

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 and only some results presented. More numerical tests are fully documented in Appendix A. For clarity’s sake, mainly low dimensional cases are described in the following, but it has been verified that the approach is equally valid for high dimensional cases (up to 8 or 9 independent variables), provided of course sufficient computational resources are available. In the next subsection the overall procedure to generate synthetic data is reviewed. Subsection 7.2 presents some results relevant to scientific applications and Subsection 7.3 provides some information about the computational requirements to implement the proposed techniques.

7.1 Overall procedure for producing synthetic data

The main technique to produce synthetic data and 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 identify the equation of the hypersurface from the points previously obtained

In the following more details about this procedure are provided with the discussion particularised for the case of binary classification and 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, an adequate number of random points in the valid range of the variables are generated. 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. The equations for producing the two classes can be summarized as follow:

$$y_1 = y + \text{noise of standard deviation } \sigma + \text{offset}$$

$$y_2 = y + \text{noise of standard deviation } \sigma - \text{offset}$$

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

In the third step, an SVM with "Gaussian Radial Basis Function kernel" is trained. The method used to find the separating hyperplane is "Sequential Minimal Optimization". 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 calculation of the boundary equations

GP Parameters	Value(s)
Population size	500
Selection method	Ranking and Tournament
Fitness function	A.I.C.
Constant range	Integers between -10 and 10
Maximum depth of trees	5
Genetic operators **(Probability)**	Crossover (45 %) Mutation (45 %) Reproduction (10 %)

In the fourth step, a mesh on the domain has to be 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 6.

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 reported in Table II.

In Appendix A, various examples are provided to illustrate the applicability and capability of the presented methodology for systems of increasing dimensionality and complexity. In next Subsection 7.2, a case at relatively high dimensionality and level of noise is discussed in detail, to illustrate the potential and main aspects of the procedure in conditions relevant to real life applications.

7.2 Effect of noise and high dimensional data

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. 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 be solved. As an example, a quite demanding example is reported in the following, for an equation involving 7 variables. The equation used to generate the data is:

$$y = x_1 \, x_2 + \sin(x_3) + \cos(x_4) - x_5 / x_6$$

It is worth mentioning that in many applications in physic and chemistry one has to deal with problem of a dimensionality not higher than 7. A total of 4000 points, 2000 per class, have been generated starting from the previous equation; more details about the synthetic data are provided in Table III. After generating the grid, training the SVM and finding the hyper-surface points, SR via GP Genetic has been applied and the following expression for the hyper-surface has been found:

$$y = 0.9 \, (x_1 \, x_2 + \sin(x_3) + \cos(x_4) - x_5 / x_6)$$

The equation identified by the method is practically the original one. The slightly different multiplicative factor in front is not to be ascribed to a weakness of the method but to the dataset provided as input, since the accuracies of both the SVM and the mathematical equation obtained are equal to 100%. Again, this example proves that, provided the surface of the boundary between the cases is sufficiently regular, the dimensionality is not an insurmountable issue provided enough computational power is available.

Table III: The function used to generate the data and the range of variables.

Steps:	Values:
Initial Function	$y = x_1 \, x_2 + \sin(x_3) + \cos(x_4) - x_5 / x_6$
Ranges of Variables	$0 < x_1 < 2$ & $1.5 < x_2 < 3$ $-2 < x_3 < 4$ & $0 < x_4 < 6$ $4 < x_5 < 12$ & $1 < x_6 < 4$
Number of Nodes for Each Class	*2000*

The numerical examples presented previously and in Appendix A include cases where the success rate of the SVM classification is close to 100%. This is an interesting situation from a scientific point of view; the SVM has learned almost perfectly the boundaries between the classes and therefore the main issue remaining consists of formulating the equations of these boundaries in a mathematical form appropriate for understanding the phenomena. If the data are such that the success rate of classification of the SVM is lower, the proposed method works well anyway, since its objective is the reformulation of the boundary equation found by the SVM. The success rate required for the SVM and the interpretation of the results is an issue, which depends on the application and the objective of the analysis, but does not impact on the validity of the developed technique. It is worth also emphasising that the task of SR in this context is not to improve the success rate of the SVM classification. The real goal consists of representing the equations of the boundary between the classes in more realistic and interpretable mathematical formulations, so that they can be used by the scientists for actual understanding (for example for comparison with theories and first principle models). To achieve this, a reasonable degradation in the success rate of classification is tolerable and typically not a major issue. In any case, with an appropriate implementation of the proposed method, typically the performance of SVM can be preserved by the final equations obtained with symbolic regression.

It is worth also mentioning that, in all the cases tested (see also Appendix A), even if the final models of the boundaries obtained by the SVM allow classifying with almost 100% accuracy, they have nothing to do with the equations generating the data. Indeed, whatever the actual formula generating the data, the model of the SVM is always of the form of equation (5). Therefore in many scientific applications, whose objective consist of understanding the physics or chemistry behind the boundaries and not simply achieving high classification rates, the SVM are not of much use except when combined with SR, as proposed in this work.

7.3 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 SR calculation 48 hours. The number of points on the grid is $16^4 * 51$; 16 for the four independent variables and 51 for the dependent one. Therefore the calculation of the grid is not a major issue since the step requiring by far most of the computational resources is the SR. On the other hand, it should be mentioned that the codes used to obtain these results had not been parallelized. In this respect, the run time to train the SVM is not a major issue, since it is typically of the order of minutes and therefore negligible compared to the other steps of the procedure. Therefore, since both the building of the grid and the Genetic Programs can be easily parallelized, reduction of the computational resources of orders of magnitude could be easily achievable.

8 Real world examples

To show the potential of the proposed methodology to attack real life problems, in this section its application to some experimental databases is reported. They have been collected in the framework of various disciplines. The first example is a typical case of a major issue in Big Physics experiments, namely Magnetic Confinement Nuclear Fusion (MCNF); the determination of the boundary between the safe and disruptive regions of the operational space. For this case all the various aspects of the proposed method are described, for the case of probabilistic SVM. Since the signals analysed are time series, particularl emphasis has also been given to the fact that this problem illustrates the potential of the approach for cases which do not satisfy the i.i.d. hypothesis. The other two consist of important examples of remote sensing in the field of atmospheric physics and for brevity sake only the main aspects of the technique are covered. The term remote sensing indicates the set of techniques aimed at obtaining information about objects without being in contact with them. These techniques can be used to monitor various aspects of the atmosphere and also the effects of human activities on the environment. One example is a case of imagery applied to the assessment of the health of vegetation. The other involves the analysis of laser backscattering signals for the detection of forests fires. For these two examples of application to remote sensing, the traditional SVM method has been implemented. The excellent results obtained in these real life applications prove the value and the flexibility of the proposed methodology.

8.1 The identification of the boundary between disruptive and safe regions of the operational space in Tokamaks.

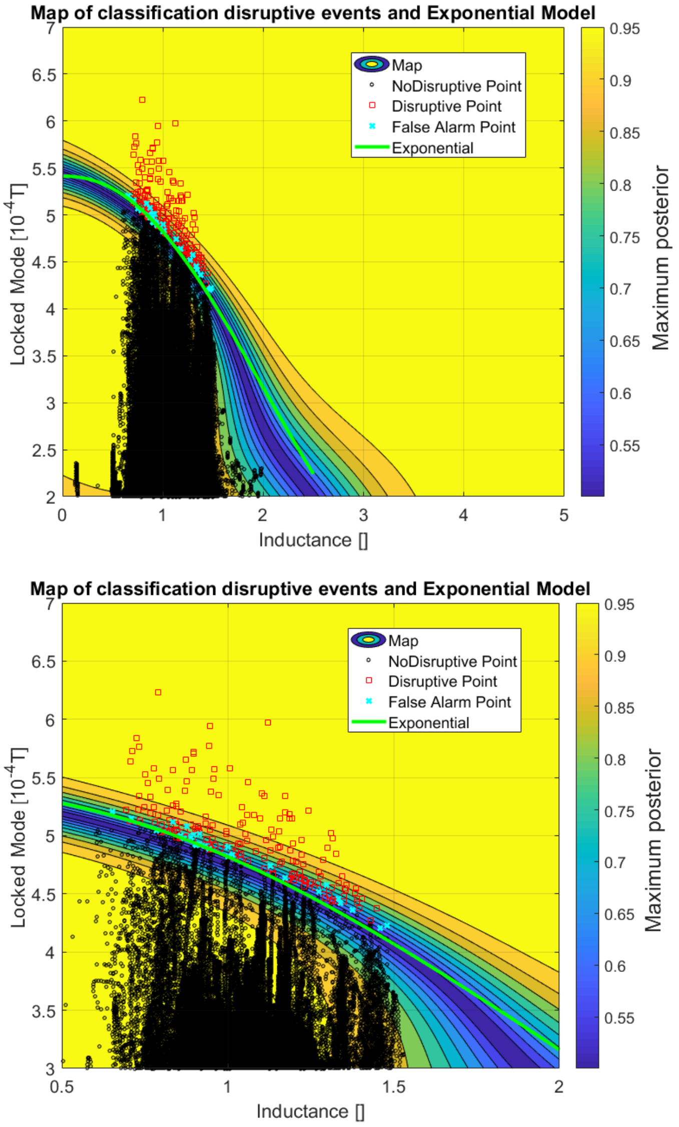

Figure 6. Top: plot of the safe and disruptive regions of the operational space in JET with the ILW. The colour code represents the posterior probability of the classifier. The black circles are all the non-disruptive shots (10 random time slices for each shot). The red squares are the data of the disruptive 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.

In the last years, collapses and their causes have become not only a major field of research but have also captured the attention of the mainstream media. From market crashes to earthquakes and structural failures in civil engineering, increasing attention is devoted to surprising and typically unexpected abrupt changes in systems, leading to catastrophic consequences. The statistical investigation of these phenomena, particularly for robust prediction, requires the development of new mathematical tools [21]. The systematic use of machine learning methods for this purpose is continuously increasing.

In Tokamaks, disruptions remain the most serious cause of collapse. Disruptions are sudden losses of confinement, leading to the abrupt quenching of the plasma with potential major risks for the structural integrity of the devices [1]. Since the potential hazard posed by disruptions increases with dimensions, the percent of disruptions allowed in the next generation of devices is quite limited. But disruptions are also a serious issue for the present largest devices. For example, they are one of the main impediments to systematic high current operation in JET [1], particularly now that the new combination of materials, Be in the main chamber and W in the divertor, renders the first wall less forgiving than in the past.

Given their potential impact on the integrity of the devices, disruptions are a subject of extensive research at present. Various methods of mitigation are being investigated, particularly massive gas injection and shatter pellets [22]. The main objective of these techniques consists of limiting the energy conducted directly to the wall by converting the highest percentage of it into radiation. On the other hand, these conversion methods have not only to be effective but also are required not to pose themselves other hazards to the machines, such as excessive increases of the eddy currents due to very fast current quenches. To reduce the strain on the devices also avoidance tactics are being considered, to undertake remedial actions and prevent the occurrence of disruptions. This is particularly important in the perspective of the final reactor, since already in the demonstrative fusion reactor unmitigated disruptions will have to be almost completely avoided and the number of mitigated ones minimised [23].

Of course, reliable prediction tools are a prerequisite to any mitigation or avoidance strategy. Unfortunately, the theoretical understanding of the causes of disruptions is not sufficient to guarantee reliable predictions. As a consequence, existing first principle models are not effective in predicting disruptions on a routine basis. Therefore, in the last decades, a lot of efforts have been devoted to developing empirical models, capable of launching an alarm when a disruption is approaching. Various generations of predictors based on machine learning tools have also been applied to JET data in the last decades. Many alternatives have been explored, ranging from Neural Networks to Self Organizing Maps and fuzzy decision trees [5-8]. Unfortunately all these different solutions are practically black boxes, who can help in practice but have not so far contribute much to the understanding of the physics behind disruptions As a results still today the major prediction tool deployed on JET is the Locked Mode Predictor based on a Threshold criterion (LMPT), which triggers mitigation actions when the signal of the locked mode amplitude reaches a certain threshold. This solution results in so called “ephemeral predictors”, i.e. systems which age very quickly and require frequent adjustments to remain effective. Indeed, in the case of LMPT, the threshold has to be adjusted quite frequently and certainly more than once per experimental campaign, not always at an optimal level [24].

To show the potential of the method proposed in this paper to find the boundary between the safe and disruptive regions of the operational space, a large database of JET, including thousands of experiments of the largest device in the world, has been analysed (see Appendix B). A systematic analysis with the CART approach has shown that, among the global quantities available in real time on JET, the locked mode and internal inductance signals are among the most relevant for disruption prediction. This confirms various manual studies performed in the past that have shown the importance of these two quantities in predicting the occurrence of disruptions not only on JET but also on other devices. Therefore, also for continuity with the past treatments, they are the two features adopted in this pilot study. The posterior probabilities have then been calculated as indicated in Section 3. The adaptive training, described in detail in Appendix C, has been performed for a whole range of threshold probabilities. It turns out that the probability value, which provides the best performance in terms of success rate, 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 to determine the most appropriate threshold probability are reported in Appendix D.

Table IV. 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 obtained at the end of the last campaign back to the entire set of data. The terms Tardy alarms, Missed alarms and Early alarms are defined in Appendix D.

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 curve level plots of the posterior probability obtained are reported in Figure 6. 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 IV, from which it is easy to appreciate the extremely good performance of the probabilistic SVM.

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:

$$y(x) = a_0 \exp(a_1 x^{a_2}) \qquad (15)$$

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

$$\begin{aligned} a_0 &= \ 5.4128 \pm 0.0031; \\ a_1 &= -0.11614 \pm 0.00085; \qquad (16) \\ a_2 &= \ 2.21 \pm 0.011; \end{aligned}$$

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 V.

Table V: The figures of merit obtained using equation (15). The terms Tardy alarms, Missed alarms and Early alarms are defined in Appendix D.

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 IV with Table V, 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 (15) is shown in light blue in Figure 6; 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 (15) represents a major simplification compared to the sum of tens of Gaussians centred on the support vectors, the model of the original SVM training. From the point of view of the physics interpretation, equation (15) 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 complements other treatments, such as the one proposed in [25], 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 . The results obtained with the proposed approach , independently from the details of the physics involved, have also important practical implications because it is clear from equation (15), and the experimental evidence of Figure 6, 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.

8.2 Botany: "wilt" database

For applying our algorithm to real-world remote sensing problems, we selected first a database related to botany named "wilt". This database was prepared by Brian Johnson from the Institute of Global Environmental strategies in Japan in 2013 and contains the results of a remote sensing study about detecting diseased trees with Qickbird imagery [26]. The data set consists of image segments, generated by segmenting the pansharpened pictures. The segments contain spectral information from the Quickbird multispectral image bands and texture information from the panchromatic (Pan) image band. In the following, the entries of this database are listed:

Class: 'w' (diseased trees), 'n' (all other land cover)
GLCM_Pan: GLCM mean texture (Pan band)
Mean_G: Mean green value
Mean_R: Mean red value
Mean_NIR: Mean NIR value
SD_Pan: Standard deviation (Pan band)

This database contains 4339 samples: 74 of them related to diseased trees and the rest related to all other land cover. The new proposed methodology has been applied to this database for finding the classification hyper-surface between the two mentioned classes. The entries have been classified first with the SVM (with the RBF kernel). The subsequent application of our technique to traditional SVM, grid plus SR, has allowed finding the following equation:

$$Mean\_G = 22.39 * Mean\_R \wedge 0.4705 \qquad (17)$$

Train Accuracy: 99.4%   Test Accuracy: 99.5 %

Since it presents a success of 99%, practically the same as the SVM, the derived equation (17) indicates that the important attributes for classifying this database are the Mean green values and the Mean red values. Figure 7 reports the entries of the database projected on the plane of these two variables, together with the hyper-surface obtained with equation (17).

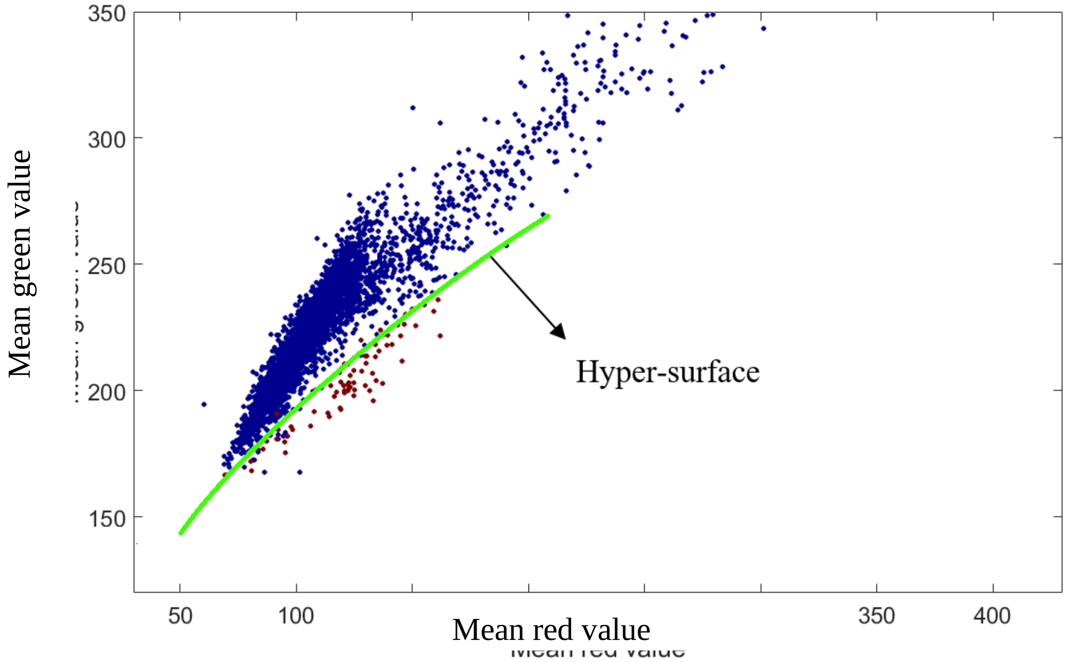

Figure 7: Distribution of data in the “wilt” database. The red points are diseased trees and the blue points indicate all other types of land cover. The black line indicates the equation obtained for the hyper-surface.

It is also worth mentioning that, to obtain the same success rate, the SVM has to utilise 1299 support vectors. Therefore the application of the proposed methodology results in a simplification of orders of magnitude in the complexity of the equation, without any significant loss in terms of classification accuracy. Moreover, the obtained formula is susceptible of comparison with models and theoretical considerations, whereas the SVM model is practically intractable from this point of view.

8.3 Remote sensing of the environment: detection of widespread smoke with LIDAR

One of these remote sensing techniques, which is gaining increasing importance, is LIDAR an acronym of Light Detection And Ranging. Lidar originated in the early 1960s, shortly after the invention of the laser, and combines laser-focused imaging with radar's ability to calculate distances by measuring the time for a signal to return. Its first deployment was in meteorology and now it is popularly used as a technology to make high-resolution maps, with applications in geomatics, archaeology, geography, geology, geomorphology, seismology, forestry, remote sensing, atmospheric physics, laser altimetry and contour mapping.

Wild fires have become a very serious problem in various parts of the world. The LIDAR technique has been successfully applied to the detection of the smoke plume emitted by wild fires, allowing the reliable survey of large areas [27- 32]. Recently, mobile compact systems have been successfully deployed in various environments. Up to now, the attention has been devoted to early detection of quite concentrated smoke plumes, characterising the first stage of fires, as soon as possible. The main operational approach consists of continuously monitoring the area to be surveyed with a suitable laser and, when a significant peak in the backscattered signal is detected, an alarm is triggered. In these applications, the backscattered signal presents strong peaks, which are detected with various techniques. In other applications, it would be interesting also to detect the non concentrated, widespread smoke, which can be the consequence of strong wind dispersion or non concentrated sources [33]. In this case, the signature of the presence of the smoke is not a strong peak in the detected power but an overall increase of large regions of the curve. Typical examples of backscattered signals for the alternatives of no smoke, strong smoke plume and widespread smoke are shown in Figure 8.

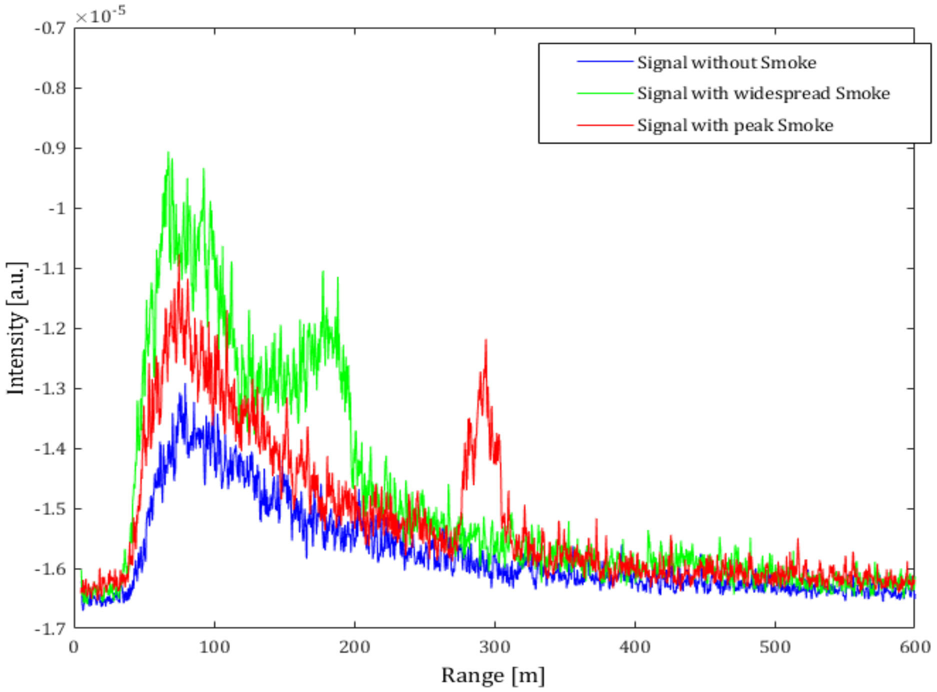

Figure 8 – Examples of LIDAR back scattered signals: a) Clear atmosphere (blue line) b) strong smoke plume (green line) c) widespread smoke (red line).

Starting from the typical Lidar equation [27], it has been decided to fit the backscattered signal intensity with a mathematical expression of the form:

$$P = \frac{K_1}{R^2} \exp\left(-2K_2 R\right) \qquad (18)$$

where K_1 and K_2 are constants and R is the range. The data of Figure 8 have been fitted with this formula. The results of the non –linear fit are:

- In case of widespread smoke:

$$P = \frac{2.648 \cdot 10^{-1}}{R^2} \cdot \exp(-1.259 \cdot 10^{-3} \cdot R) \qquad (19)$$

- Clear atmosphere:

$$P = \frac{1.734 \cdot 10^{-1}}{R^2} \cdot \exp(-1.171 \cdot 10^{-3} \cdot R) \qquad (20)$$

The results of the fit, equations (19) and (20), indicate quite clearly that the parameter K_2 are very similar for both the case of widespread smoke and clear atmosphere. On the other hand, there is a clear difference, of the order of 25% in the constants K_1. This is expected since K_1 includes the effect of the coefficient β, which indeed quantifies the backscattering properties of the atmosphere [27, 30].

Since the attempt to identify the presence of widespread smoke is a quite pioneering application of the LIDAR technique, it is important not only to be able to discriminate between the two situations but also to provide models for the interpretation of the physics. In particular, the identification of the boundary in the space of the parameters K_1 and K_2 for the two cases is considered an essential piece of information for comparison with theories. The proposed methodology has therefore been applied to a quite substantial database:

Total number of data = 521
number of non-smoke data = 312
number of widespread smoke data = 209

number of train data (~80%) = 431
number of test data (~20%) = 90

For the SVM, a radial basis functions kernel has been used. The best equation found is:

$$K_1 = 0.1083 * \sin (15.61 * K_2 \wedge 2) + 0.1083 * \cos (1.5941 * K_2 \wedge 0.264) \qquad (21)$$

Train Accuracy: 89.33 %   Test Accuracy: 91.11 %

The equation of the boundary between clear atmosphere and widespread smoke, in the space of the parameters K_1 and $K_{2,}$ is shown in Figure 9.

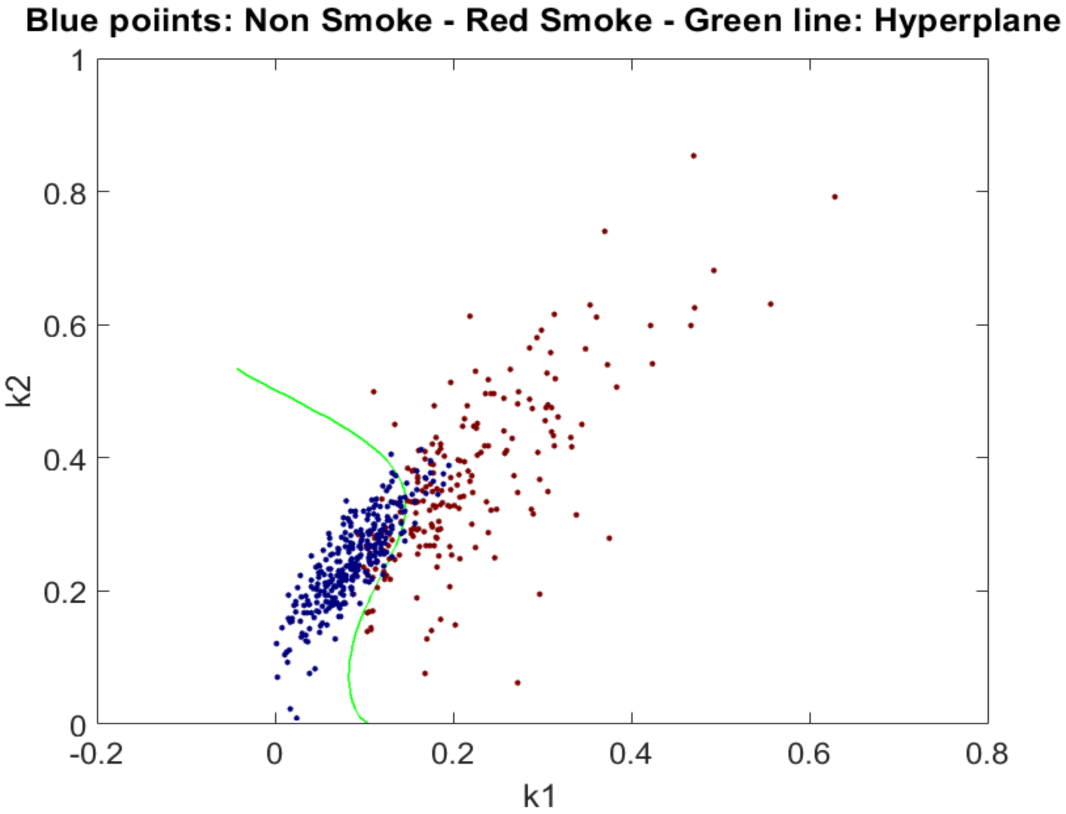

Figure 9 – Equation (21), describing the boundary between the boundary between the cases of clear atmosphere and widespread smoke, in the space of the parameters K$_1$ and K$_2$.

To understand the importance of the results obtained, it should also be considered that the model of the SVM consists of 154 support vectors. Therefore the level of simplification obtained with equation (21) is substantial. Moreover, also in this case the formalism of the SVM provides an equation of the boundary between the two classes which has no relation with the relevant physics.

9 Conclusions

An original methodology has been devised to obtain the equation of the boundary between two classes, using an array of machine learning tools, including almost all the main machine learning techniques available. With the proposed approach, the power of machine learning tools is combined with the realism, physics fidelity and interpretability of equations expressed in the usual formalism of typical scientific theories. In particular, the noise-based ensemble of CART trees has proved essential in identifying the most important features to include in the analysis in an efficient way, taking into account the problem of the noise from the first step of the treatment. The choice of SVM ensures that their structural stability, their capability to maximize the safety margins in the classification, is fully retained in the final result. On the other hand, symbolic regression via genetic programming allows achieving very good physics fidelity and finding a good trade-off between accuracy of the classification and complexity of the final equations of the boundary. Therefore the models obtained with the proposed methodology are able to better support fundamental scientific activities such as testing of mathematical theories, evaluation of confidence intervals, scaling, extrapolations and experimental design. It is also worth mentioning that “a priori” information can be exploited in order to steer the solutions towards mathematical expressions, which reflect the actual dynamics of the phenomena under study. This can be achieved for example by selecting properly the basis functions or by constraining the structure of the trees.

Given the fact that the objectives of the approach are realism and interpretability, a reasonable reduction of the classification performance is not a major issue and can be tolerated. It is also true that symbolic regression via genetic programming can reproduce the accuracy of the classification by the SVM, provided a sufficiently high number of nodes is allowed in the final solution. Indeed, in all the tested numerical cases, no reduction of the classification performance has been found provided the necessary complexity of the SR tress has been implemented.

It is also worth emphasizing one more time that the proposed procedure is fully coherent in the treatment of the error bars of the measurements, a very important aspect in the perspective of the application of the developed tools in scientific domains. Indeed from the the noise-based ensemble and the choice of the σ in the RBS kernel of the SVM, to the use of the GD as the fitness function of the SR, the effect of the uncertainties in the measurements can be fully taken into account in all the steps of the procedure. Even the non-linear fitting, typically required after the mathematical form of the equations has been obtained with SR, can be performed using the GD.

Another important aspect of the methodology is the adaptive training of the SVM. In may applications in the natural sciences the i.i.d. hypothesis is far from being satisfied and therefore traditional training approaches are conceptually unsatisfactory. The consequence of the violation of the i.i.d. assumption is typically the ephemeral character of the models. With the approach proposed in the paper, this problem is completely remedied and the evolution of the equations describing the boundary between the classes can be followed in detail.

The numerical tests shown have proved the effectiveness of the proposed technique to identify the real equation of the boundary between classes even in relatively high dimensions, provided the shape of the boundary is a sufficiently regular surface. Again, this seems to be fully adequate since, in the majority of the scientific applications, the boundaries between the various classes are quite regular functions. This has been confirmed by the application of the technique to experimental databases of different scientific disciplines.

On the other hand, the method is susceptible of various improvements. First of all, the technique should be extended to other machine learning tools such a neural networks (the only major thread of machine learning not included in the present version of the methodology). More fundamentally, the approach is now limited to identifying the mathematical expressions of boundaries which can be expressed as functions. It is a topic of future investigations to apply the method to the investigation of more complex boundaries (for example multiply connected hypersurfaces). Moreover, the task of regression, and not only classification, should also be tackled [34]. Also applications to various aspects of tomography inversion are envisaged [35, 36]. Another important theoretical aspect is the extension of the approach to cases affected by different statistics of the noise (and not only the usual Gaussian). In this respect, advances in information geometry, with the formulation of geodesic distances valid for other noise statistics, are considered the right direction of future work. From the computational point of view, the heaviest step of the proposed methodology is SR via GP. It is clear that this part of the method is highly parallelizable; therefore much progress is expected in the reduction of running time by parallel implementation of SR via GP.

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] J. Wesson, "*Tokamaks*". Oxford : Clarendon Press Oxford, 2004. Third edition.

[2] G. A. Rattá, et al. Nuclear Fusion. 50 (2010) 025005 (10pp).

[3] J. Vega, et al. 2014. "*Adaptive High Learning Rate Probabilistic Disruption Predictors from Scratch for the Next Generation of Tokamaks*" Accepted for publication in Nuclear Fusion.

[4] A.Murari, et al Nucl. Fusion **49** (2009) 055028 (11pp)

[5] B. Cannas et al JET Nucl. Fusion 53 093023 doi:10.1088/0029-5515/53/9/093023.

[6] A.Murari et al, Nucl. Fusion **53** (2013) 033006 (9pp)

[7] J. Vega, A. Murari, G. Vagliasindi, G. A. Rattá and JET-EFDA Contributors. Nuclear Fusion. 49 (2009) 085023 (11pp)

[8] P.Gaudio et al. Plasma Phys. Control. Fusion 56 (2014) 114002.

[9] V.Vapnik "*The Nature of Statistical Learning Theory*" Springer Science & Business Media, 2013, ISBN 1475724403, 9781475724400

[9] García, S., Fernández, A., Luengo, J. et al. Soft Comput (2009) 13: 959. https://doi.org/10.1007/s00500-008-0392-y

[10] A Vellido, JD Martín-Guerrero, PJG Lisboa - ESANN, 2012 - elen.ucl.ac.be

[11] Leo Breiman, Jerome Friedman, Charles J. Stone, R.A. Olshen "*Classification and regression trees*" Taylor & Francis, 1984

[12] M.Lungaroni et al On the Potential of Ruled-Based Machine Learning for Disruption Prediction on JET Submitted ot Fusion Engineering and Design

[13] V. Vapnik "*The nature of statistical learning theory*" Information Science and Statistics Springer 2000

[14] 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.

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

[16] M.Schmid, H.Lipson, Science, Vol 324, April 2009

[17] Koza J.R., *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA (1992).

[18] Kenneth P. Burnham, David R. Anderson (2002), *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer. (2nd ed)

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

[20] Shun-ichi Amari and Hiroshi Nagaoka, "Methods of information geometry" Translations of Mathematical Monographs, Vol 191, Oxford University Press, 2000

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

[22] S.Meitner et al Fusion Science and Technology Journal Volume 72, 2017 Issue 3

[23] 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

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

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

[26] Johnson, B., Tateishi, R., Hoan, N., 2013. A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. International Journal of Remote Sensing, 34 (20), 6969-6982.

[27] G. Fiocco and L. D. Smullin, Nature, 199, 1275 (1963).

[28] F. Andreucci, M. Arbolino, A study on forest fire automatic detection system, Il Nuovo Cimento, 16, 1, 35 (1993).

[29] C. Bellecci, M. Francucci, P. Gaudio, M. Gelfusa, S. Martellucci, M. Richetta, T. Lo Feudo, Appl. Phys. B 87, 373 (2007).

[30] C. Bellecci, P. Gaudio, M. Gelfusa, T. Lo Feudo, A. Murari, M. Richetta, L. De Leo, In-cell measurements of smoke backscattering coefficients using a CO2 laser system for application to lidar-dial forest fire detection, Optical Engineering, 49 (12), 124302 (2010).

[31] J. Vega, A. Murari, S. González and JET-EFDA contributors, "A universal support vector machines based method for automatic event location in waveforms and video-movies: applications to massive nuclear fusion databases", Review of Scientific Instruments, vol. 81 (2), p. 023505 (2010).

[32] M. Gelfusa, P. Gaudio, A. Malizia, A. Murari, J. Vega, M. Richetta, S. Gonzalez," UMEL: A new regression tool to identify measurement peaks in LIDAR/DIAL systems for environmental physics applications", Review Scientific Instr., 85, 063112 (2014)

[33] M. Gelfusa, A. Malizia, A. Murari, S. Parracino, M. Lungaroni, E.Peluso, J. Vega, L. DeLeo, C. Perrimezzi and P. Gaudio: "First attempts at measuring widespread smoke with a mobile lidar system" submitted to IEEE Xplore

[34] A.Murari et al Nucl. Fusion **55** (2015) 073009 (14pp) doi:10.1088/0029-5515/55/7/073009

[35] L Marrelli, P Martin, A Murari, G Spizzo "Total radiation losses and emissivity profiles in RFX" Nuclear fusion 38 (5), 1998, 649

[36] P Martin et al "Soft x-ray and bolometric tomography in RFX" Review of scientific instruments 68 (2), 1997, 1256-1260

Appendix A Numerical Tests for SR via GP to obtain realistic boundary equations

The procedure described in Section 6 has been subjected to a systematic series of numerical tests. A significant set of these tests is reported in this Appendix.

A.1 Examples for two independent variables

Example 1

As a first test, a purely arithmetic function has been tested. The function and ranges of the variables are:

$$y = x_1 + x_2 - x_1 \times x_2 \qquad -1 < x_1 < 1 \qquad 1 < x_2 < 2$$

After carrying out the six-step procedure described in Section 6, the following expression has been obtained:

$$y = 1.011\,(\,x_1 + x_2 - x_1 \times x_2\,)$$

SR via GP converges on a final expression that is in excellent agreement with the initial function describing the boundary between the two classes. This is particularly true since such a good approximation has been obtained without the non-linear fitting, normally the last step of the SR method.

Example 2

As a second test, a more complex function comprising exponential, arithmetic, and power operators has been assumed for the boundary between the two classes. The function and ranges of the variables are:

$$y = exp\,(\,(\,x_1 \times x_2\,)^{0.5}\,) \qquad 0 < x_1 < 1 \qquad 1 < x_2 < 3$$

After carrying out the six-step procedure in Section 6, the following expression has been obtained:

$$y = 0.974\,exp\,(\,(\,x_1 \times x_2\,)^{0.5}\,)$$

Again 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.

Example 3

As the third test, a more complex function comprising trigonometric and arithmetic operators has been defined and 4% classification noise was added to the database. The function and ranges for the variables are:

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

After carrying out the six-step procedure in Section 6, the following expression has been obtained:

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

Again 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 A1 presents the results of this example in pictorial form.

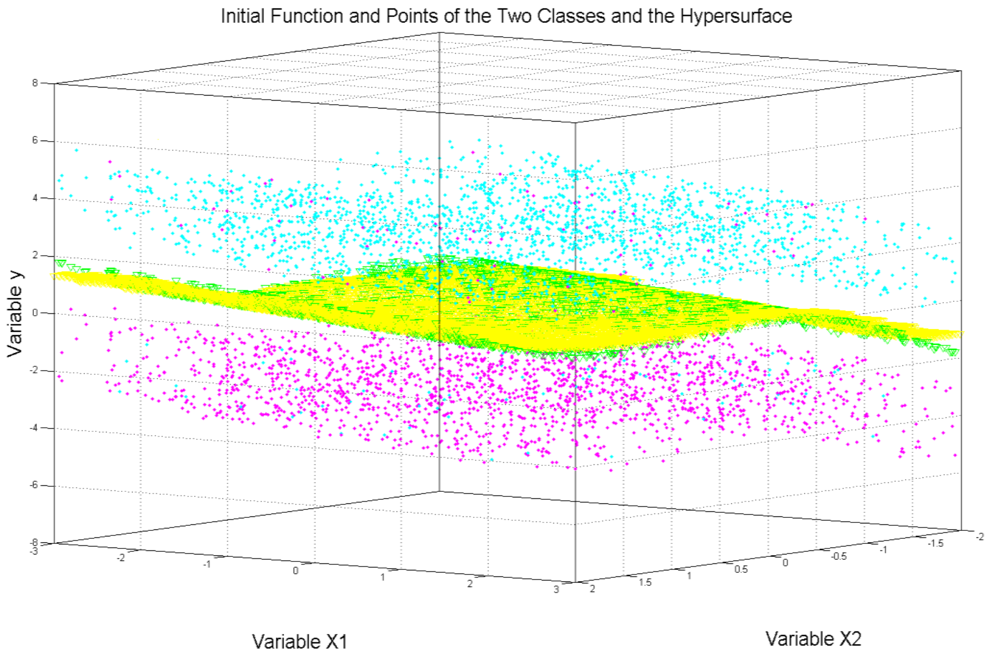

Figure A1: Points and surfaces of example 3 with two independent variables. Green rectangles are points generated from the initial function, Cyan points are the points belonging to the first class, Magenta points are the points belonging to the second class, and the Yellow surface identifies the hyper-surface obtained with the SR via GP.

A.2 Examples for three independent variables

Some examples considering equations with three independent variables are reported in this section.

Example 1

As a first test, a function comprising only arithmetic operators has been defined. The function and ranges for the variables are:

Initial Defined Function: $y = x_1 - x_2 + x_3$

Range of Variables: $1 < x_1 < 2$ $3 < x_2 < 5$ $0 < x_3 < 1$

The final function obtained from the hypersurface points is:

$$y = 1.002 (x_1 - x_2 + x_3)$$

Example 2

As a second test, a function comprising trigonometric and arithmetic operators has been defined. The function and ranges for the variables are:

Initial Defined Function: $y = x_1 + \sin (x_2 \times x_3)$

Range of Variables: $1 < x_1 < 2$ $3 < x_2 < 5$ $0 < x_3 < 1$

The final function obtained from the hypersurface points is:

$$y= 0.98 (x_1 + \sin (x_2 \times x_3))$$

Again these results confirm the great potential of the approach. Almost exactly the original function can be obtained already at the stage of SR. With additional rounding off of the results or application of non linear fitting, exactly the original functions can easily be recovered.

A.3 Example for four independent variables

In this subsection, we describe the results of the application of the SVM-GP methodology to a more complex and noisy database. A five-dimensional synthetic database has been generated with the characteristics described in Table AI.

Table AI 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	$-1.5 < x_1 < 1.5$ & $-2 < x_2 < 2$ $0 < x_3 < 2$ & $2 < x_4 < 4$
Number of Nodes for Each Class	*2000*
Thickness of the data's bulk	*3*
Offset	*10% of y domain*
Classification Noise	~ *4%*

The procedure for 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 AII.

Table AII: The success rates of 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:

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

The obtained equation is in good agreement with the initial function. 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 AIII (to be compared with the results reported in Table AII).

Table AIII : The success rates obtained for the train and test data for the classification of the synthetic database with the expression obtained via SR

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

The comparison of the accuracies obtained via SVM and with our proposed technique allows concluding that the SVM-GP approach has excellent performance, even for more complex databases and in higher dimensions, in interpreting the SVM hyper-plane as a hyper-surface equation.

Appendix B Database of JET with a metallic wall

All experiments in JET campaigns C29 to C31 have been considered. After proper cleaning and validation of the DB, overall 187 disruptive and 1020 non disruptive shots are included, unless differently specified. JET database with the ILW has been used to implement the methodology described in this paper. 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.

Appendix C Adaptive training

The theory of most if not all machine learning tools is based on the so called i.i.d. assumption; the examples are meant to be independent and identically distributed. In practice this means that the pdf generating the data is the same and the examples are drawn independently from it. This assumption is clearly violated in most situations and experiments in the natural sciences. Certainly JET experiments are very different from one another and evolve in an historical way. To overcome this problem, an adaptive training from scratch has been devised. 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 already been developed.

Appendix D Performance of the probabilistic SVM described in Section 8.1 for various choices of the triggering window: database of JET with the ILW.

Table D1. Main figures of merit of the probabilistic SVM 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

Table D2. Main figures of merit of the probabilistic SVM 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

Table D3. Main figures of merit of the probabilistic SVM 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