

ELM Regime Classification by Conformal Prediction on an Information Manifold

Aqsa Shabbir, Member, *IEEE*, Geert Verdoolaege, Member, *IEEE*, Jesús Vega and Andrea Murari

Abstract—Characterisation and control of plasma instabilities known as edge-localised modes (ELMs) is crucial for the operation of fusion reactors. Recently, machine learning methods have demonstrated good potential in making useful inferences from stochastic fusion data sets. However, traditional classification methods do not offer an inherent estimate of the goodness of their prediction. In this work, a distance-based conformal predictor classifier integrated with a geometric-probabilistic framework is presented. A first benefit of the approach lies in its comprehensive treatment of highly stochastic fusion data sets, by modeling the measurements with probability distributions in a metric space. This enables calculation of a natural distance measure between probability distributions: the Rao geodesic distance. Secondly, the predictions are accompanied with estimates of their accuracy and reliability. The method is applied to the classification of regimes characterized by different types of edge-localized modes based on measurements of global parameters and their error bars. This yields promising success rates and outperforms state-of-the-art automatic techniques for recognizing ELM signatures. The estimates of goodness of the predictions increase the confidence of classification by ELM experts, while allowing more reliable decisions regarding plasma control and at the same time increasing the robustness of the control system.

Index Terms— Conformal predictors, edge-localized modes, geodesic distance, information manifold.

I. INTRODUCTION

HIGH confinement or H-mode plasmas in tokamaks are usually characterized by cyclic instabilities near the plasma edge, referred to as *edge-localized modes* or ELMs. ELMs result in a sudden exhaust of particles and energy but are nonetheless advantageous for attaining stationary plasma conditions as they result in impurity and helium ash expulsion. With ELMs as the basis for distinction, H-mode plasmas can be roughly categorized into three types: ELM-free H-mode, H-mode with small ELMs and H-mode

with large or type I ELMs. The H-mode with relatively large low-frequency type I ELMs has become the reference plasma scenario for ITER, and beyond. Characteristics of ELMs, their control and comprehensive physical understanding are crucial for ITER and next step fusion devices. On one hand the beneficial properties of ELMs, in terms of enhanced edge particle transport are well recognized, on the other hand, there has been concern that on future large devices giant ELM bursts could damage divertor and first wall surfaces or disrupt internal transport barriers [1]. This has motivated intensive research for accomplishing effective ELM control and mitigation. Optimization of control and mitigation mechanisms and enhancement of the physical understanding necessitates the discrimination of different observed classes of ELMs. In contrast to the existing mostly phenomenological categorizations of ELM types, this work is aimed at developing a data-driven methodology for automatic classification and discrimination of ELMs.

Recently, machine learning and pattern recognition techniques have shown substantial potential in data-driven studies of fusion plasmas by extracting useful patterns of interest from fusion data [2]-[5]. This yields an important tool for real-time plasma control, e.g. in ITER, in order to maintain good plasma equilibrium or control certain types of instabilities. Moreover, a data-driven study of the primary physical variables that determine the confinement regimes and instabilities, such as ELMs, can improve substantially the understanding of the governing physical mechanisms.

The objective of the present work is twofold. First, we wish to contribute to the discrimination of diverse ELM behavior by presenting an effective methodology for quantitative distinction between ELM types. Second, for practical purposes we aim to contribute to the dependability and robustness of control strategies by providing a discriminator for ELM types equipped with estimates of reliability and accurateness. We present an automated classification system for ELM types and apply our method to classify regimes with small and type I ELMs. The system, currently, makes use of a standard set of global plasma and engineering variables related to plasma confinement. The focus of the current work is to obtain better classification rates compared to existing classifiers and thus the obtained success rates can be further optimized by using more informative plasma and engineering parameters.

The act of classification is fundamentally related to the occurrence of *clustering structure* in the data space, where each cluster of measurement points corresponds to a certain plasma phenomenon, such as a specific type of ELMs. Hence our method falls within the domain of *pattern recognition methods*, with the clusters constituting a pattern in the data

This work was supported by the European Commission and carried out within the framework of the Erasmus Mundus International Doctoral College in Fusion Science and Engineering (FUSION-DC).

A.Shabbir is with the Department of Applied Physics, Ghent University, 9000 Ghent, Belgium and Max-Planck Institute for Plasma Physics, 85748 Garching, Germany (e-mail: aqsa.shabbir@ugent.be).

G.Verdoolaege is with the Department of Applied Physics, Ghent University, 9000 Ghent, Belgium and the Laboratory for Plasma Physics, Royal Military Academy (LPP-ERM/KMS), 1000 Brussels, Belgium.

J.Vega is with Laboratorio Nacional de Fusión, CIEMAT, 28040 Madrid, Spain.

A. Murari is with Consorzio RFX (CNR, ENEA, INFN, Università di Padova, Acciaierie Venete SpA), 35127 Padua. Italy.

space, reflecting an important aspect of the physics of the plasma.

The classification system proposed in this paper is integrated with the probabilistic data representation framework presented earlier by Verdoolaege *et al.* [3]-[4]. The primary motivation for this framework is the substantial uncertainty that frequently characterizes the measurements of plasma quantities, which may contain both a stochastic and systematic component. The main factors contributing to stochastic uncertainty are hardware noise and plasma fluctuations, since these can usually not be modelled tractably in a deterministic way. The proposed framework takes into account the statistical error bars or, more generally, the stochastic features of the data, by modelling the data with suitable probability distributions. In order to characterize the data patterns, such as clusters, in the associated probabilistic space, a similarity measure between probability distributions is required. The mathematical field of information geometry provides an appropriate similarity measure between probability density functions (PDFs), which are interpreted as points on a Riemannian differentiable manifold, or *information manifold* [6]-[8]. The PDF parameters provide a coordinate system on the manifold and the Rao geodesic distance (GD) serves as a natural similarity measure between PDFs. The classifier, which then operates in this information space, is based on *conformal predictors* (CPs), first described by Vovk *et al.* [9] and Saunders *et al.* [10]. Conformal prediction offers various advantages over the traditional machine learning methods (MLMs). Most noteworthy, they provide information about their own accuracy and reliability with the only assumption of randomness of the data samples. Also known as the *iid* hypothesis, the randomness assumption implies that all training samples are independent of each other and are identically distributed according to the same (but unknown) distribution [11]. Unlike traditional MLMs, CPs do not enforce a rigid separation between learning and prediction, but learn dynamically alongside making predictions. Furthermore, they do not require prior probabilities as the Bayes classifier and also have the ability to detect ambiguities in the classification task, i.e., when a unique class cannot be assigned to a new example.

In this paper a computationally efficient nearest-neighbor CP coupled with the geometric-probabilistic data representation framework is deployed for classification of H-mode plasma regimes into H-mode with small ELMs and H-mode with type I ELMs. The proposed technique is compared with a discriminant analysis classifier and a nearest-neighbor classifier, which are well-established state-of-the-art MLMs. The presented technique not only yields higher classification accuracy, but also returns a quantitative estimate of the prediction's accuracy and reliability, which traditional MLMs do not provide. Furthermore, the classification performance is calculated for both the geodesic distance geometry of the data and the conventional Euclidean distance. The geodesic distance improves the classification performance, establishing itself as a natural similarity measure between probability distributions lying on an information manifold.

The outline of the paper is as follows. In Section II we discuss the modalities of our proposed geometric-probabilistic framework and the details of the approach. Section III

discusses the application of conformal predictors to ELM identification in relation to our modeling framework. Section IV presents the experimental setup, visualization and classification results and their analysis. Section V concludes the paper.

II. A GEOMETRIC-PROBABILISTIC PATTERN RECOGNITION FRAMEWORK

A. The geometry of probability distributions

The Fisher information can be regarded as a metric tensor (Fisher-Rao metric) on an information manifold, which is a Riemannian differentiable manifold formed by a family of PDFs, such as the Gaussian family [3]-[4]. Once the metric is known, geodesic equations can be established and solved, allowing for the calculation of the geodesic (shortest-path) distances on the manifold [7]-[8]. Given a probability model $p(\mathbf{x}|\boldsymbol{\theta})$ for a vector-valued variable \mathbf{x} , labelled by an m -dimensional parameter vector $\boldsymbol{\theta}$, the components of the Fisher information matrix $g_{\mu\nu}$ are defined through the relations

$$g_{\mu\nu}(\boldsymbol{\theta}) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^\mu \partial \theta^\nu} \ln p(\mathbf{x}|\boldsymbol{\theta}) \right] \quad (1)$$

$\mu, \nu = 1, \dots, m.$

B. The geometry of the univariate Gaussian distribution

In this paper we model the data using a simple univariate Gaussian model. The Fisher-Rao metric for the Gaussian distribution, parameterized by its mean μ and standard deviation σ , can be given via the quadratic line element [12]:

$$ds^2 = \frac{1}{\sigma^2} d\mu^2 + \frac{2}{\sigma^2} d\sigma^2 \quad (2)$$

A closed-form expression exists for the GD, permitting a fast evaluation. Indeed, for two univariate Gaussian distributions $p_1(x|\mu_1, \sigma_1)$ and $p_2(x|\mu_2, \sigma_2)$, parameterized by their mean μ_i and standard deviation σ_i ($i = 1, 2$), the GD is given by [12]

$$GD(p_1||p_2) = \sqrt{2} \ln \frac{1+\delta}{1-\delta} \quad (3)$$

$$\delta \equiv \frac{[(\mu_1 - \mu_2)^2 + 2(\sigma_1 - \sigma_2)^2]^{\frac{1}{2}}}{[(\mu_1 - \mu_2)^2 + 2(\sigma_1 + \sigma_2)^2]^{\frac{1}{2}}} \quad (4)$$

A convenient Gaussian geometric model is provided by the Poincaré half-plane, which is represented in Fig. 1(a). The horizontal axis corresponds to the mean μ of the Gaussian distribution, while on the positive part of the vertical axis the standard deviation σ is represented. Every point in this half-plane corresponds to a unique Gaussian and the geodesics between two points are half-circles as well as half-lines ending on the horizontal axis, the latter connecting distributions that differ only in their standard deviation (not drawn). The distance between points along one of these curves in the

Poincaré half-plane is the same as the actual geodesic distance between the points. The evolution of the distribution along an example geodesic is shown in Fig. 1(b).

Finally, in the case of multiple independent Gaussian variables it is easy to prove that the squared GD between two sets of products of distributions is given by the sum of the squared GDs between corresponding individual distributions [12].

III. CONFORMAL PREDICTORS

In classification systems, each observation (or sample) is expressed as an ordered pair (x_i, y_i) , where x_i is a feature vector (i.e. the set of parameters that characterize the sample i) and y_i is the class label of the observation i , where the set of labels is finite and usually small. Given a data set of N samples, a conventional MLM uses a subset of the data set, which it designates as a training set for determining the prediction rule. Then follows a testing phase, wherein a subset of the dataset is used for determining the goodness of the prediction rule. Conformal predictors obviate the need of a distinct training and testing phase, which is the premise of classical machine learning methodologies. CPs offer so-called blended learning and prediction, as they learn and predict at the same time, continuously improving their performance as they carry out each prediction and discover how accurate the prediction was. Samples that get classified are added to a hypothetical “bag” of samples and participate in the classification of the next incoming samples.

CPs estimate the goodness of their prediction by means of two figures of merit: *confidence* and *credibility*. Confidence gauges the reliability of the prediction, while credibility is an indicator of how representative the training set is for the new sample that is to be classified. New confidence values are obtained at each classification, taking into account both the previous samples that have been classified and all possible labels for the current one. For classifying each incoming sample, CPs evaluate how different the current sample is from each cluster (class) within the bag samples by determining a “nonconformity score” for the current sample with respect to each cluster (class) within the bag samples. In this work, a nearest-neighbor scheme is used for determining the nonconformity score. Essentially, the nonconformity score for the current sample is provided by its distance to its nearest neighbors for both classes, amongst the bag samples.

Specifically, the nonconformity score α_i of a given sample i is calculated as

$$\alpha_i = \frac{d_{i-SL}}{d_{i-DL}} \quad (5)$$

Where d_{i-SL} is the distance to sample i 's nearest neighbor in the bag with the same label and d_{i-DL} is the distance to sample i 's nearest neighbor in the bag with a different label.

The nonconformity score for sample i is computed with respect to both classes, assuming membership of sample i of each of the classes $j = 1, 2$ in turn. By doing this for each sample, a ranking can be determined of the nonconformity scores. Then, for each class j a p -value is calculated based on this ranking, namely:

$$p_j = \frac{\#\{i=1, \dots, M | \alpha_i \geq \alpha_q\}}{M} \quad (6)$$

Here, M is the number of bag samples, i.e. the samples that have already been classified, and α_q is the nonconformity score for the sample that is to be currently classified. The p -value is essentially the fraction of bag samples that are at least as different as the current sample. The current sample is assigned to the class with the largest corresponding p -value. The largest p -value itself is referred to as the credibility, while the complement of the other p -value is the confidence of the classification task:

$$\text{Credibility} = \max(p_j), \quad j = 1, 2 \quad (7)$$

$$\text{Confidence} = 1 - \min(p_j), \quad j = 1, 2 \quad (8)$$

The smaller p -value is essentially the probability of the prediction being in error and thus the probability of correctness of the current prediction is automatically quantified by the confidence.

In case where the p -value assigned to each class is the same, CP deems that a unique class cannot be assigned to the current sample. It refrains from making any (possibly incorrect) decision and separates the current sample whilst labelling it ambiguous.

Despite the numerous advantages offered by CPs, the method can become computationally expensive and thus infeasible in real time, for very large data sets. This is a direct consequence of the dynamic learning capabilities possessed by CP. This limitation is overcome by deploying a computationally efficient variant of CP: inductive conformal predictors (ICPs) [11]. ICPs offer a compromise between dynamic learning and computational time, without causing degradation of the classification performance. ICPs divide the dataset into two sets: the *proper training* set and the *calibration* set. The proper training set, similar to a conventional MLM, is used for computing the decision rule once, which is dynamically improved as each sample from the calibration set (a pseudo test set) is classified. In this work, ICPs as well as a theoretically pure form of CPs known as transductive conformal predictors (TCPs) are each applied. TCPs in contrast to ICPs require a minimalistic proper training data set and at the least one sample per class suffices.

As a similarity measure in calculating the nonconformity scores, we first considered the Euclidean distance between the sample's feature vector and that of the bag samples. Then we compared its performance to that of the GD, this time treating the features as quantities with an error bar, hence Gaussian distributions.

IV. ELM IDENTIFICATION

A. Physics picture of ELM types

The physical mechanisms of the different observed classes of ELMs are complex. As a result, no unified first principles theory describing ELMs exists. Type I ELMs mainly seem to be driven by the steep pressure gradient, whereas small ELMs appear to be controlled by the absolute value of the edge temperature along with steep pressure gradient. This suggests that small ELMs are linked to resistive MHD phenomena whereas type I ELMs are associated with ideal MHD [13]-[15]. Considerable progress has been made in ELM modelling activity [16]-[17], with the peeling-ballooning model appearing as the leading candidate for explaining the trigger for the ELMs. This model builds on the two instability sources near the plasma edge namely, current and pressure gradients. It outlines a pseudo-triangular operating diagram for ELMs in the space of the ballooning pressure gradient and the normalized edge current.

Currently, type I ELMs and small ELMs are primarily distinguished by their response to increased heating power. The ELM repetition frequency for type I ELMs increases with increasing power and decreases for small ELMs.

An alternate way of distinguishing between the two classes is to compare temperatures and densities at the pedestal top. In [18] it is shown that type I ELMs are clustered around a hyperbola of constant, high pedestal pressure. This constant corresponds to the theoretically predicted onset of pressure driven, ideal MHD ballooning mode instability. Small ELMs appear to occur below a critical pedestal temperature $T_{e,crit}$ which tends to increase with the toroidal magnetic field. Further, on the $n_{edge} - T_{edge}$ diagram [18] they are seen as two clusters: one at low $T_{e,ped}$ and high $n_{e,ped}$ and the other at high $T_{e,ped}$ and low $n_{e,ped}$.

A crucial distinction between type I and small ELMs is their effect on plasma confinement. Type I ELMs H-modes have superior overall plasma confinement but the ELM size poses serious concerns for future fusion machines. On the contrary, the size of small ELMs offers no concern for the machine operation but the energy confinement time is 10-30% below that in type I H-mode [14]-[15].

In order to predict ELM behavior in next step fusion devices and ensure operation in the desirable ELM regimes, development of an automated discrimination scheme for ELMs is required, constituting the starting point of this work.

B. ITPA database

In this work for ELM regime identification we employed measurements from the International Tokamak Physics Activity (ITPA) Global H-mode Confinement Database (DB3, version 13f), henceforth referred to as the “ITPA database” [19]-[20]. The ITPA database contains more than 10,000 validated measurements of various global plasma and engineering variables at one or several time instants during discharges in 19 tokamaks. The data have been used extensively for determining scaling laws for the energy confinement time, mainly as a function of a set of eight plasma and engineering parameters: plasma current, vacuum toroidal magnetic field, total power loss from the plasma, central line-averaged electron density, plasma major radius,

plasma minor radius, elongation and effective atomic mass. We have used the same eight global variables to discriminate between type I and small ELMs. Specifically, all database entries with a confinement mode labelled HGELM and HGELMH were considered to belong to the H-mode region with type I ELMs and all database entries tagged HSELM and HSELMH were regarded as belonging to the H-mode region with small ELMs. For current work, the database entries have been normalized to bring all variables in proportion with one another prior to subsequent operations.

It should be noted that classification of ELM characteristics based on global non-time-resolved data is a considerable challenge. Indeed, in addition to the information contained in the global time-averaged values of the plasma parameters, space-resolved measurements, near the plasma boundary, of the plasma density and temperature could easily improve the recognition rates. Similarly, estimates of changes in the thermal and fast particle energy content per ELM burst and measurements of ELM frequency obtained from time traces of plasma quantities, such as the D_α radiation, can also considerably improve the predictive capacity of the method. However, in the present work we did not yet take into account these additional sources of information, although our method is perfectly able to incorporate and treat these data.

The ITPA database lists typical error estimates of measurements for the various plasma and engineering variables. This represents very limited information on the probability distribution underlying each quantity. Nevertheless, effective utilization of this limited information proves beneficial. In this work it is assumed that the error bars pertain to a statistical uncertainty in the data, specifically that they represent a single standard deviation. According to the principle of maximum entropy, the underlying probability distribution is Gaussian with mean the measurement itself and standard deviation the error bar. Also, it is supposed that for stationary plasma conditions, all variables are statistically independent and so the joint distribution factorizes. This means that the joint distribution for the eight variables mentioned above is assumed to be just the product of the individual univariate Gaussian distributions. Clearly, this is a strong assumption and it is imposed here mainly for keeping the calculations tractable. It is noteworthy that our formalism has no difficulties with the heterogeneous sources of the measurements, coming from different tokamaks and possibly with different error bars for essentially the same quantities. The reason is that the error estimates are automatically embedded in the probabilistic data description.

The number of samples from each tokamak belonging to the H-mode class with ELMs is given in Table I. Further, the numbers of small and type I ELM samples per machine are also listed.

C. Visualization

Visualization of high-dimensional data sets through a projection in the two-dimensional Euclidean plane is a useful tool for enabling plasma physicists to gain knowledge about the internal structure of the data and relationships in it. Its goal is to amplify human cognition and provide an intuitive insight into the possible interactions and relationships in complex and

frequently large data sets [21]. Hence, visualization of the data within the region of the operational space corresponding to H-mode with ELMs can be very useful because it can potentially yield enhanced insight in the configuration of the operational space. It can convey important information regarding the conditions, under which specific plasma regimes occur, as well as the “distance” of the current plasma conditions from a certain desired or undesirable regime. Visualization of the operational space is not a straightforward task as the information is not normally directly available, since the number of variables labeling the operational space is often greater than two. Hence the dimensionality of the data space is higher than two, preventing a simple plot of the data in a two-dimensional diagram. Moreover, in our framework each measurement is represented by a Gaussian probability distribution with a mean and an error bar. This distribution cannot be represented by a point in a Euclidean space but naturally lies on a curved Riemannian manifold. Therefore, data visualization is a natural starting point in distinguishing between regions of different ELMs, which essentially are found in neighboring or overlapping regions of the operational space.

In this work, visualization of the high-dimensional and/or probabilistic (non-Euclidean) data is obtained by projection of the data onto a two-dimensional Euclidean plane. To do this, we use metric multidimensional scaling (MDS), which is a well-regarded information visualization technique [22] and is widely used in perceptual mapping. MDS provides a two-dimensional mapping of the ITPA database, which originally spans eight dimensions (16 in case the standard deviation of each measurement is counted as an extra parameter). In order to calculate the distance in the original high-dimensional data space, we use the GD in the case when the probabilistic representation of the data is taken into account.

A projection using MDS is shown in Fig. 2 for the entire ITPA database. Certainly, the visual map is an approximation of the original configuration, but nevertheless MDS yields a projection of points in the Euclidean space with least distortion of all pairwise distances; i.e. the mapping is approximately isometric. Hence, the real value of the projection lies in the relative position of the points with respect to each other and in contrast to usual scientific visualizations the coordinate axes are less significant. Further, a visual map from a subset of the data is also plotted in Fig. 2. These are ASDEX, AUG, JET and DIII-D, as these machines are the major contributors to the ELM data in the ITPA database. Visualizations in Fig. 2 incorporate the measurement uncertainty as MDS uses the GD between Gaussian product distributions. It can be readily noted from Fig. 2 that there is a considerable overlap between the ELM classes, rendering the classification task a veritable challenge (although it should always be remembered that the visualization is a projection, inevitably resulting in information loss). Further it can be seen in Fig. 2 (c) and 2(d) that the data of ASDEX and JET roughly conform to two clusters. This distribution is due to different levels of plasma current and toroidal fields in the machines. Also, it can be observed from visual inspection of Fig. 2 (f) that the data of AUG is heavily unbalanced with very few samples from the H-mode region with small ELMs. Despite these constraints imposed by the data set, our classification

scheme is able to attain a relatively good separation between the two classes.

D. Classification via conformal prediction

The experiments were performed for 20%, 50% and 70% of the total data being treated as a proper training set, followed by the dynamic learning for the remaining data. Proper training data were selected at random from the entire database, while ensuring the same balance with respect to the class variable as was present in the original data, i.e. if, say, the original data contained 70% samples from class 1 and 30% samples belonged to class 2, then the samples which constitute the training set also maintain the same ratio with respect to class label. The results are given in Table II. Transductive conformal prediction is also carried out for reference. This is shown in Table II as the entry which uses 0.01% of the total data as the initial proper training data. The GD is the distance measure of choice used to calculate the nonconformity scores, for the results presented in Table II. Similar experiments were conducted using the Euclidean distance, operating on the measurement values without consideration of the error bars. These results are presented in Table III. Table II and Table III each report the success rate (SR) for classification, which is the average over the two classes for correct predictions made as a percentage of total predictions. Similarly, the error rate (ER) is provided, which is the average over the two classes for the incorrect predictions made as a percentage of total predictions. Also listed is the ambiguity (AM), i.e. the ratio of the number of samples for which a prediction could not be made, for the total number of samples in the data set. The last two columns for each table contain the average values of confidence and credibility for the predictions made. The SRs achieved with the GD and Euclidean similarity measures are also illustrated in Fig. 3.

The most noteworthy outcome is that the GD gives a superior performance in contrast to the Euclidean distance, both in terms of success rate and the average confidence level of the predictions. This establishes that exploiting the information content residing in measurement uncertainty is important for identifying ELM types.

Furthermore, the CPU time (in seconds) for the CP classification obeys a linear law:

$$t = 1.30n + 48 \quad (9)$$

where n is the number of samples which constitute the calibration set, i.e. are dynamically classified. ICPs provide a significant reduction in computational time, as they effectively reduce the size of the calibration set without introducing degradation in success rates.

E. Comparative analysis

In previous works, discriminant analysis has been used for ELM identification [23]. To allow a homogeneous comparison between CPs and other well-established MLMs we perform the classification of H-mode with small and type I ELMs

using linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and a 1-nearest-neighbor classifier. 50% of the data is used as training data, thus enabling a direct comparison with the entry with 50% of the data used as proper training data in Tables II and III. Balance with respect to the class variable is kept intact in the randomly selected training data. Each experiment is repeated 10 times with different random training sets and thus each mentioned result is in fact an average over 10 replications. The average SRs for classification alongside their standard deviation (STD) are given in Table IV.

It can be readily seen that CPs provide a significantly higher success rate in contrast to well-established MLMs, in addition to providing an estimate of the classification accuracy (confidence) and reliability (credibility).

F. Validation

We further empirically validated the performance of our classification scheme and the obtained results using N -fold stratified cross-validation. This is an established model validation technique, since for an optimal choice of N , it reduces the bias in the prediction output while combating variance and yet being computationally feasible. The mechanism behind cross-validation is illustrated in Fig. 4. It operates by dividing the available data into roughly N equal parts and then iteratively training and testing the classification scheme using $N-1$ parts for training and the remaining one part for testing. Hence each sample in the data set gets eventually used for training and testing. We performed all our experiments using 10-fold cross-validation. The obtained results are given in Tables V and VI.

The success rates are higher for each classification scheme, compared to the results in the previous section. However, the important observation is that CP consistently performs better than the other techniques. Furthermore, again the GD measure gives the better performance as compared to the Euclidean distance. Fig. 5 illustrates the success rates for each classification scheme. The results are shown for 10-fold cross-validation and also for 50% of the total data being used for training (from section 4.3). Superior performance of CP coupled with GD can be readily noted.

G. Performance for individual machines

Finally, we provide success rates for classification of ELM regions for individual machines. Results are given in Table VII and were obtained using a 10-fold cross-validated inductive conformal predictor with the geodesic distance. Class-wise success rates are also given for each machine, where the two classes are H-mode region with small ELMs, denoted by ‘S’ and H-mode region with type I ELMs, denoted by ‘Type I’.

For analyzing the results given in Table VII, the following characteristics of data need to be considered:

- Class imbalance: a two-class data set is considered imbalanced (or skewed) when one of the classes is heavily under-represented in comparison to the other class.
- Dispersion of data: degree to which the data points within a cluster are dispersed over the feature space.

Class imbalance for each machine is listed in Table VIII, while the statistics of dispersion within the class are given in Table IX. Dispersion is quantized by computing the mean distance to the nearest neighbor within the class. The larger the mean distance to the nearest neighbor, the larger is the spread within the class. Furthermore, higher is the standard deviation in the distance to the nearest neighbor for the samples, the lower is the likelihood of occurrence of localized clusters of a certain class. The distance to nearest neighbor is based on the GD between the probability distributions.

The data from JET are balanced and also have smallest within-class dispersion of all the machines. This can also be seen by visual inspection of Fig. 2(d), where localized clusters of each class can be observed despite the limitations of the projected space. Hence classification performance for data from JET is high on the whole and also for each class individually. The data from ASDEX are not just partially unbalanced but also suffer from large within-class dispersion. This is also verified by the visual projection in Fig. 2(c). As a consequence, classification performance is lower than that of other machines. The data from AUG are highly unbalanced in the favor of class type I, i.e. H-mode region with type I ELMs. As a result, the classification performance for class type I is very high and that of class S, i.e. H-mode region with small ELMs, is considerably lower. However, the high success rate for class type I outweighs the other effects, making the overall average success rate for this machine the highest. The success rate for class S for DIII-D is the lowest amongst all machines. Once again, this can be attributed to a partial imbalance of the class towards class type I and a higher dispersion within class S as compared to class type I.

H. Practical implications

The most significant contribution of this work is the development of an alternative methodology for classification of ELM types. It is noteworthy that the relative performance of the proposed method in contrast to other techniques is more important than the absolute value of the success rate. This is so because the absolute success rates are a stronger indicator of the quality of the chosen features and their discriminatory power for the problem at hand, rather than the goodness of the classification scheme. Incorporating additional features such as pedestal parameters, ELM frequency etc. is likely to significantly improve success rates and forms a part of the authors’ ongoing work. Further, more accurate estimates of error bars and more precise information regarding the distribution of the uncertainties could be very useful for optimizing the classification performance. Suitability and advantages of the developed method have been demonstrated using the ITPA database, despite the limited information

offered by it on the underlying probability distribution of the predictors. Having established the merits of the technique it can be rightfully argued that the method can be deployed as one of the layers in an embedded multi-layer classifier for ELM types.

V. CONCLUSIONS

We have presented an approach that offers a new perspective to the discrimination of ELM types, as an addition to the existing predominant phenomenological categorizations. The presented approach conducts pattern recognition using global plasma data while consistently taking into account uncertainties: first uncertainty of the data themselves and then of the classification results. This is important in order to increase the reliability of classifiers for resolving the underlying physics and for plasma control decisions.

The proposed distance-based conformal predictor classifier integrated with the geometric-probabilistic framework provides for an automated classifier for ELM types with high success rates and a figure of its own merit: confidence and credibility. Furthermore, it possesses dynamic learning capability and a mechanism for detection of ambiguities, which is advantageous over incorrect classification. The method is generic and can be applied to other problems in nuclear fusion, such as disruption prediction. In addition, it is exportable to other application domains in signal and image processing. The method can also potentially help in quantifying the change in the behavior of ELMs in response to control and mitigation strategies.

ACKNOWLEDGMENT

This work was supported by the European Commission and carried out within the framework of the Erasmus Mundus International Doctoral College in Fusion Science and Engineering (FUSION-DC).

REFERENCES

- [1] E.J. Doyle, W.A. Houlberg, Y. Kamada, V. Mukhovatov, T.H. Osborne, A. Polevoi, G. Bateman, J.W. Connor, J.G. Cordey, T. Fujita, X. Garbet, T.S. Hahm, L.D. Horton, A.E. Hubbard, F. Imbeaux, F. Jenko, J.E. Kinsey, Y. Kishimoto, J. Li, T.C. Luce, Y. Martin, M. Ossipenko, V. Parail, A. Peeters, T.L. Rhodes, J.E. Rice, C.M. Roach, V. Rozhansky, F. Ryter, G. Saibene, R. Sartori, A.C.C. Sips, J.A. Snipes, M. Sugihara, E.J. Synakowski, H. Takenaga, T. Takizuka, K. Thomsen, M.R. Wade, H.R. Wilson, ITPA Transport Physics Topical Group, ITPA Confinement Database and Modelling Topical Group and ITPA Pedestal and Edge Topical Group, "Chapter2: Plasma confinement and transport," *Nucl. Fusion*, vol.47, no.6, pp. 18-127, 2007.
- [2] A. Murari, J. Vega, D. Mazon, G.A. Rattà, J. Svensson, S. Palazzo, G. Vagliasindi, P. Arena, C. Boulbe, B. Faugeras, L. Fortuna, D. Moreau and JET-EFDA Contributors, "Innovative signal processing and data analysis methods on JET for control in the perspective of next-step devices," *Nucl. Fusion*, vol. 50, no.5, art. No. 055005, 2010.
- [3] G. Verdoolaege, G. Karagounis, A. Murari, J. Vega, G. Van Oost, and JET-EFDA contributors, "Modelling fusion data in probabilistic metric spaces: applications to the identification of confinement regimes and plasma disruptions," *Fusion Sci. Technol.*, vol.62, no.2, pp. 356-365, 2010.
- [4] G. Verdoolaege, G. Karagounis, M. Tendler and G. Van Oost, "Pattern recognition in probability spaces for visualization and identification of plasma confinement regimes and confinement time scaling," *Plasma Phys. Control. Fusion*, vol.54, no.12, art. No.124006, 2012.
- [5] A. Shabbir, G. Verdoolaege, G. Van Oost, J. Noterdaeme, and JET-EFDA Contributors, "Discrimination and visualization of ELM types based on a probabilistic description of inter-ELM waiting times," in *Proc. 41st EPS Conf. Plasma Phys.*, P1.001, Berlin, Germany, 2014.
- [6] S. Amari, and H. Nagaoka, "Methods of information geometry," *Trans. Math. Monographs*, vol. 191. AMS, 2000.
- [7] G. Verdoolaege and P. Scheunders, "On the geometry of multivariate generalized Gaussian models," *J. Math. Imaging Vis.*, vol. 43, no. 3, pp. 180-193, 2011.
- [8] G. Verdoolaege and P. Scheunders, "Geodesics on the manifold of multivariate generalized Gaussian distributions with an application to multicomponent texture discrimination," *Int. J. Comput. Vision*, vol. 95, pp. 265-286, 2011.
- [9] V. Vovk, A. Gammerman, and C. Saunders, "Machine learning applications of algorithmic randomness," in *Proc. 16th Int. Conf. Mach. Learning*, pp. 444-453, San Francisco, CA, 1999.
- [10] C. Saunders, A. Gammerman, and V. Vovk, "Transduction with confidence and credibility," In *Proc. 16th Int. Joint Conf. Artif. Intell.*, vol. 2, pp. 722-726, 1999.
- [11] V. Vovk, A. Gammerman and G. Shafer, "Algorithmic Learning in a Random World," Springer, 2005.
- [12] J. Burbea and C. Rao, "Entropy differential metric, distance and divergence measures in probability spaces: a unified approach," *J. Multivariate Anal.*, vol.12, no.4, pp.575-596, 1982.
- [13] H. Zohm, "The physics of edge localized modes (ELMs) and their role in power and particle exhaust," *Plasma Phys. Control. Fusion*, vol.38, pp. 1213-1223, 1996.
- [14] J.W. Connor, "Edge-localized modes – physics and theory" *Plasma Phys. Control. Fusion*, vol.40, no.5, pp. 531, 1997.
- [15] M. Bécoulet, G. Huysmans, Y. Sarazin, X. Garbet, Ph. Ghendrih, F. Rimini, E. Joffrin, X. Litaudon, P. Monier-Garbet, J.-M. Ané, P. Thomas, A. Grosman, V. Parail, H. Wilson, P. Lomas, P. de Vries, K.-D. Zastrow, G. F. Matthews, J. Lonnroth, S. Gerasimov, S. Sharapov, M. Gryaznevich, G. Counsell, A. Kirk, M. Valovic, R. Buttery, A. Loarte, G. Saibene, R. Sartori, A. Leonard, P. Snyder, L. L. Lao, P. Gohil, T. E. Evans, R. A. Moyer, Y. Kamada, A. Chankin, N. Oyama, T. Hatae, N. Asakura, O. Tudisco, E. Giovannozzi, F. Crisanti, C. P. Perez, H. R. Koslowski, T. Eich, A. Sips, L. Horton, A. Hermann, P. Lang, J. Stober, W. Suttrop, P. Beyer, S. Saarelma and Contributors to JET-EFDA Workprogramme, "Edge localized mode physics and operational aspects in tokamaks," *Plasma Phys. Control. Fusion*, vol.45, no.12A, pp. A93-A113, 2003.
- [16] H.R. Wilson, P.B. Snyder, G.T.A. Huysmans and R.L. Miller, "Numerical studies of edge-localized instabilities in tokamaks," *Phys. Plasmas*, vol. 9, no.4, 2002.
- [17] P.B. Snyder, H.R. Wilson, J.R. Ferron, L.L. Lao, A.W. Leonard, D. Mossessian, M. Murakami, T.H. Osborne, A.D. Turnbull, and X.Q. Xu, "ELMs and constraints on the H-mode pedestal: peeling-ballooning stability calculation and comparison with experiment," *Nucl. Fusion*, vol.44, no.2, pp. 320, 2004.
- [18] ITER Physics Expert Group on Confinement and Transport, ITER Physics Expert Group on Confinement Modelling and Database and ITER Physics Basis Editors, "Chapter 2: Plasma confinement and transport," *Nucl. Fusion*, vol.39, no.12, pp.2175-2249, 1999.
- [19] International Tokamak Physics Activity H-mode Database Working Group: <http://efdasql.ipp.mpg.de/hmodepublic> (2007).
- [20] D.C. McDonald, J.G. Cordey, K. Thomsen, O.J.W.F. Kardaun, J.A. Snipes, M. Greenwald, L. Sugiyama, F. Ryter, A. Kus, J. Stober, J.C. DeBoo, C.C. Petty, G. Bracco, M. Romanelli, Z. Cui, Y. Liu, Y. Miura, K. Shinohara, K. Tsuzuki, Y. Kamada, T. Takizuka, H. Urano, M. Valovic, R. Akers, C. Brickley, A. Sykes, M.J. Walsh, S.M. Kaye, C. Bush, D. Hogewei, Y.R. Martin, A. Cote, G. Pacher, J. Ongena, F. Imbeaux, G.T. Hoang, S. Lebedev, A. Chudnovskiy and V. Leonov, "Recent progress on the development and analysis of the ITPA global H-mode confinement database," *Nucl. Fusion*, vol.47, no.3, pp. 147-174, 2007.
- [21] A. Shabbir, G. Verdoolaege, O.J.W.F. Kardaun, J.-M. Noterdaeme and JET-EFDA Contributors, "Visualization of the operational space of edge-localized modes through low-dimensional embedding of probability distributions," *Rev. Sci. Instrum.*, vol. 85, issue 11, article no. 11E819, 2014.
- [22] T.F. Cox and M.A.A. Cox, "Multidimensional Scaling," Chapman and Hall, 2001.
- [23] O.J.W.F. Kardaun, "Classical methods of statistics," Springer, 2005.

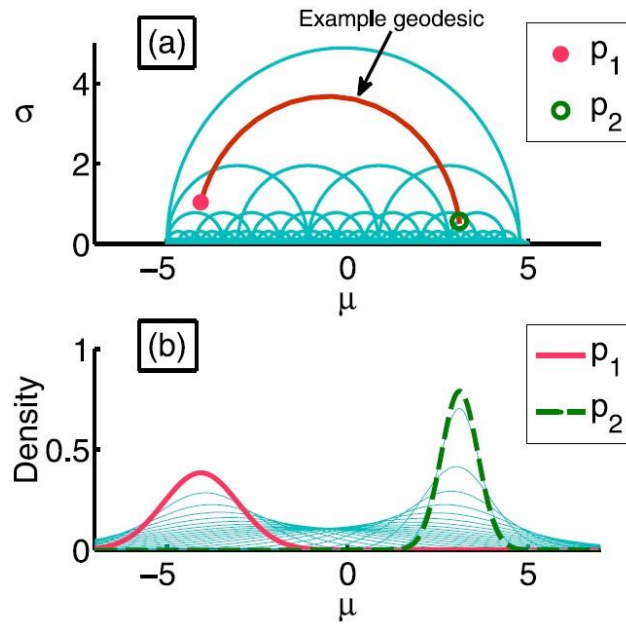


Fig. 1. (a) Illustration of the Poincaré half-plane with several half-circle geodesics, one of them between the points p_1 and p_2 . b) Probability densities corresponding to the points p_1 and p_2 indicated in (a). The densities associated with some intermediate points on the geodesic between p_1 and p_2 are also drawn.

• Small ELMs ★ Type I ELMs

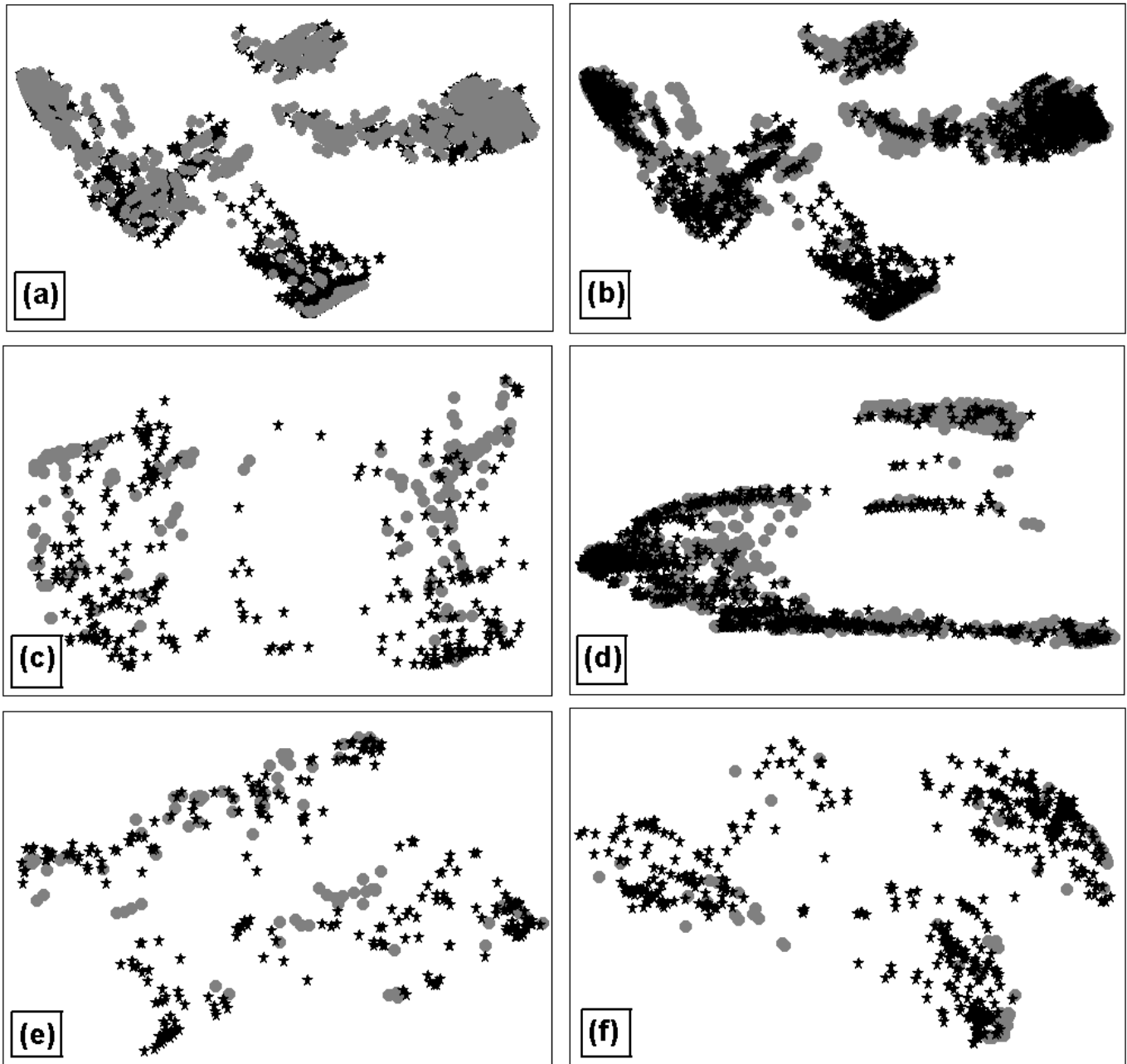


Fig. 2. Two-dimensional projections using MDS with indicated small ELM and type I ELM clusters. (a) ELM data from the entire ITPA database with small ELMs on top. (b) ELM data from the entire ITPA database with type I ELMs on top. (c) ELM data from ASDEX. (d) ELM data from DIII-D. (f) ELM data from AUG.

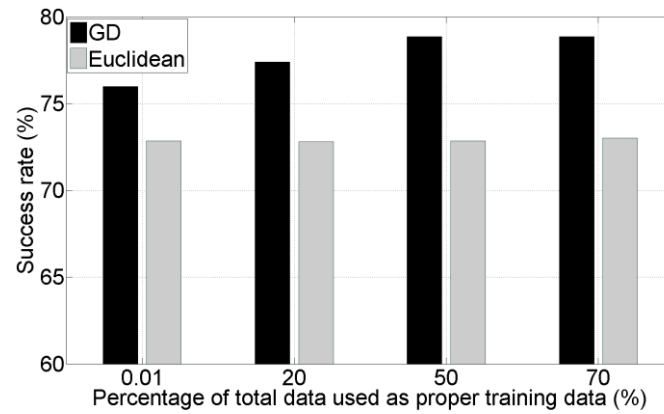


Fig. 3. Success rates (%) for CP

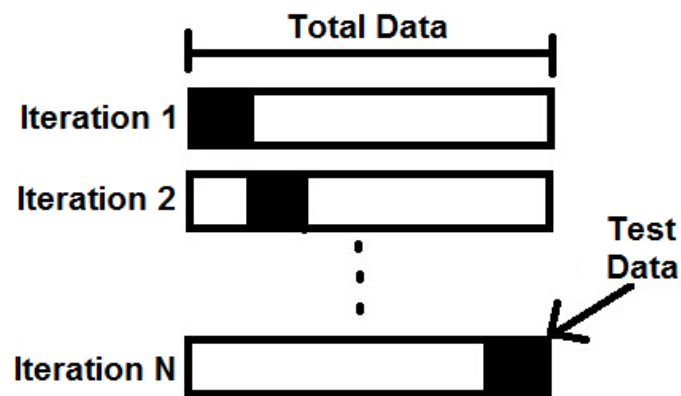


Fig. 4. Prototype for N-fold cross-validation

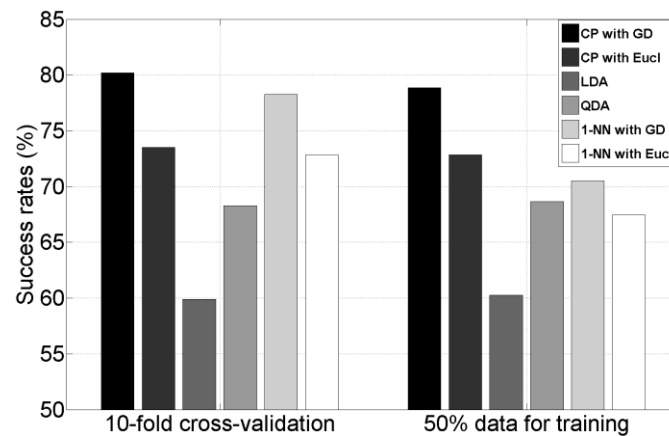


Fig. 5. Comparative success rates (%) for different classification schemes. Results are shown for both 10-fold cross-validation and random sampling of training data, where 50% of the total data is selected for training.

TABLE I

TOTAL NUMBER OF SAMPLES FROM EACH TOKAMAK IN THE ITPA DATABASE BELONGING TO THE H-MODE REGION WITH ELMS. THE NUMBER OF SAMPLES PER CLASS, I.E. SMALL AND TYPE I ELMS, IS ALSO GIVEN.

Machine	Total samples	Type I ELMS	Small ELMS
ASDEX	445	287	158
AUG	583	498	85
CMOD	46	0	46
COMPASS	26	13	13
DIII-D	343	249	94
JET	1780	980	800
JFT-2M	76	0	76
JT60-U	89	35	54
PBXM	80	19	61
PDX	117	48	69
TCV	15	15	0
TFTR	99	5	94
TDEV	10	0	10
START	9	0	9
MAST	12	0	12
NSTX	6	0	6

TABLE II

SUCCESS RATES (SR), ERROR RATES (ER), AMBIGUITIES (AM) AND AVERAGE CONFIDENCE (CO) AND CREDIBILITY (CR) FOR CLASSIFICATION OF ELM TYPES USING VARIOUS SIZES OF THE PROPER TRAINING DATA SET AND WITH THE GEODESIC DISTANCE AS THE SIMILARITY MEASURE.

Initializing training data (%)	SR (%)	ER (%)	AM (%)	CO (%)	CR (%)
0.01	75.98	23.89	0.108	92.3	57.7
20	77.39	22.57	0.034	92.5	55.4
50	78.85	21.06	0.108	92.7	49.6
70	78.58	21.31	0.089	93.0	45.6

TABLE III

SIMILAR TO TABLE II, BUT WITH THE EUCLIDEAN DISTANCE AS THE SIMILARITY MEASURE.

Initializing training data (%)	SR (%)	ER (%)	AM (%)	CO (%)	CR (%)
0.01	72.85	27.04	0.11	90.5	56.8
20	72.80	27.20	0.00	89.6	55.7
50	72.84	27.16	0.00	89.6	51.3
70	73.01	26.99	0.00	89.8	48.1

TABLE IV

AVERAGE SUCCESS RATES (SR) AND CORRESPONDING STANDARD DEVIATIONS (STD) FOR THE CLASSIFICATION OF REGIMES WITH DIFFERENT ELM TYPES BASED ON LINEAR DISCRIMINANT ANALYSIS (LDA), QUADRATIC DISCRIMINANT ANALYSIS (QDA), 1-NEAREST NEIGHBOR (1-NN) USING GD AND 1-NN USING THE EUCLIDEAN DISTANCE.

Classifier		SR (%)	STD (%)
LDA		60.26	0.39
QDA		68.65	0.33
1-NN	GD	70.50	0.97
1-NN	Euclidean	67.47	0.48

TABLE V

SUCCESS RATES (SR), ERROR RATES (ER), AMBIGUITIES (AM) AND AVERAGE CONFIDENCE (CO) AND CREDIBILITY (CR) FOR CLASSIFICATION OF ELM REGIMES USING 10-FOLD CROSS-VALIDATION WITH THE GD AND EUCLIDEAN DISTANCE AS A SIMILARITY MEASURE.

Classifier	Distance Measure	SR (%)	ER (%)	AM (%)	CO (%)	CR (%)
CP (ICP)	GD	80.19	19.69	0.134	94.1	43.5
	Euclidean	73.50	26.48	0.027	90.2	48.1

Table VI

SUCCESS RATES (SR) AND CORRESPONDING STANDARD DEVIATION (STD) FOR THE CLASSIFICATION OF ELMY REGIMES WITH 10-FOLD CROSS-VALIDATION USING LINEAR DISCRIMINANT ANALYSIS (LDA), QUADRATIC DISCRIMINANT ANALYSIS (QDA), 1-NEAREST NEIGHBOR (1-NN) USING THE EUCLIDEAN DISTANCE AND FINALLY 1-NN USING THE GD.

Classifier		SR (%)	STD (%)
LDA		59.89	0.92
QDA		68.28	0.81
1-NN	GD	78.27	0.88
1-NN	Euclidean	72.85	0.67

Table VII

SUCCESS RATES (SR), ERROR RATES (ER), AMBIGUITIES (AM), AVERAGES CONFIDENCE (CO) AND AVERAGE CREDIBILITY (CR) FOR ELM REGIME CLASSIFICATION USING ITPA DATA FROM JET, ASDEX, AUG AND DIII-D, BASED ON A 10-FOLD CROSS-VALIDATED INDUCTIVE CONFORMAL PREDICTOR.

Machine	Class	SR(%) by class	SR (%)	ER (%)	AM (%)	CO (%)	CR (%)
JET	S	76.76	78.32	21.34	0.34	94.1	43.5
	Type I	79.88					
ASDEX	S	64.23	70.46	28.46	1.14	91.2	44.2
	Type I	76.70					
AUG	S	69.35	82.86	16.49	0.53	97.3	40.1
	Type I	96.37					
DIII-D	S	61.10	74.33	25.67	0.00	93.3	40.6
	Type I	87.55					

Table VIII

NUMBER OF SAMPLES FROM EACH CLASS FOR EACH MACHINE, DETERMINING THE RESPECTIVE CLASS BALANCE.

Machine	Class	No. of samples per class	Ratio of class S to class Type I	Class balance
JET	S	800	45:55	Balanced
	Type I	980		
ASDEX	S	158	35:65	Partially Unbalanced
	Type I	287		
AUG	S	85	15:85	Unbalanced
	Type I	498		
DIII-D	S	94	27:73	Partially Unbalanced
	Type I	249		

Table IX

MEASURE OF DISPERSION WITHIN EACH CLASS FOR EACH MACHINE.

Machine	Class	Mean distance to NN	Standard deviation
JET	S	1.75	2.33
	Type I	1.48	2.02
ASDEX	S	9.13	3.01
	Type I	9.83	2.93
AUG	S	12.78	2.79
	Type I	13.08	2.27
DIII-D	S	4.56	3.20
	Type I	2.69	2.00