EFDA–JET–CP(08)04/03

N. Duro, R. Dormido, J. Vega, S. Dormido-Canto, G. Farias, J. Sánchez,
H. Vargas, A. Murari and JET EFDA contributors

Automated Recognition System for ELM Classification in JET

Automated Recognition System for ELM Classification in JET

N. Duro[1], R. Dormido[1], J. Vega[2], S. Dormido-Canto[1], G. Farias[1], J. Sánchez[1],
H. Vargas[1], A. Murari[3] and JET EFDA contributors*

JET-EFDA, Culham Science Centre, OX14 3DB, Abingdon, UK

[1]Dpto. de Informática y Automática - UNED. C/ Juan del Rosal 16, 28040 Madrid, Spain
[2]Asociación EURATOM/CIEMAT para Fusión. Avd. Complutense 22, 28040 Madrid, Spain
[3]Consorzio RFX-Associazione EURATOM ENEA per la Fusione. I-35127 Padua, Italy
** See annex of M.L. Watkins et al, "Overview of JET Results ",*
(Proc. 21 st IAEA Fusion Energy Conference, Chengdu, China (2006)).

Preprint of Paper to be submitted for publication in Proceedings of the
25th Symposium on Fusion Technology, Rostock, Germany
(15th September 2008 - 19th September 2008)

ABSTRACT
Edge Localized Modes (ELMs) are instabilities occurring in the edge of H-mode plasmas. Considerable efforts are being devoted to understanding the physics behind this non-linear phenomenon. A first characterization of ELMs is usually their identification as Type I or Type III. An automated pattern recognition system has been developed in JET for off-line ELM recognition and classification. The empirical method presented in this paper analyzes each individual ELM instead of starting from a temporal segment containing many ELM bursts. The ELM recognition and isolation is carried out using three signals: D_α, line integrated electron density and stored diamagnetic energy. A reduced set of characteristics (such as diamagnetic energy drop, ELM period or D_α shape) has been extracted to build supervised and unsupervised learning systems for classification purposes. The former are based on Support Vector Machines (SVM). The latter have been developed with hierarchical and k-means clustering methods. The success rate of the classification systems is about 98% for a database of almost 300 ELMs.

1. INTRODUCTION
Plasma instabilities should be successfully controlled in order to produce fusion energy efficiently and without compromising the material boundary. Edge localised modes (ELMs) are one of these instabilities that are not fully known and further theoretical and experimental analysis are required. In order to advance in the study of the physics behind ELMs, data-driven methods seem to be powerful techniques to extract knowledge from the experimental signals without assuming any kind of hypothesis. This knowledge could be combined with theoretical models for both exploratory and confirmatory analysis.

This article develops a data-driven approach for the characterization and automatic classification [1] of ELMs as Type I or Type III [2]. To this end, three steps are accomplished. The first one is to identify, isolate and extract individual ELMs from JET signals (in the present approach, each individual ELM is analysed instead of starting from a temporal segment containing many ELM bursts). Although the physical basis for this assumption is not established, we introduce it as working hypothesis. The second step is a feature extraction process to represent the ELMs with a minimum set of relevant characteristics. In the third step, three classifications methods (supervised and unsupervised) have been applied to classify the ELMs. All computations have been performed by developing several software tools based on MATLAB.

Section 2 describes the identification and feature extraction; section 3 explains the three classification methods that we have used; section 4 shows results and, finally, section 5 presents some conclusions.

2. IDENTIFICATION AND FEATURE EXTRACTION OF ELMS
As it was mentioned above, this article presents a method that analyzes each individual ELM. The ELM recognition and isolation is carried out using three signals: stored diamagnetic energy (corresponding to the JET signal MG3F/WPD), line integrated electron density (JET signal KG1V/

LID4) and D_α (JET signal S3AD/ AD34). ELMs are recognized by an abrupt change in the diamagnetic energy and a simultaneous drop in the line integrated electron density. As a consequence of the ELM instability, a typical peaked shape appears in the D_α (Fig.1). At present, the isolation procedure of ELMs is carried out in a manual way.

Once individual ELMs are identified, the feature extraction process must be accomplished. It consists of extracting features or attributes that are of distinctive nature. The set of attributes allow achieving a proper representation of the ELMs that will be used in the classification process. Typically in classification processes, the selection of both the number of features and the specific ones that best represent the objects is problem dependent. It requires specific knowledge about the system to classify. In practice, a larger than necessary number of feature candidates is generated and then the 'best' of them are adopted.

In the case of ELMs, features have been chosen with the aim of incorporating some knowledge of the experts of ELMs. The diamagnetic energy has been the reference signal and, in particular, the drop in this magnitude has been selected. In the first approximation, Type I ELMs seem to show larger drops in the diamagnetic energy than Type III ELMs. Another attribute is the ELM period. The repetition frequency is often used as a characterisation parameter of ELM types. Again in the first approximation, high/low rates are assigned to Type III/Type I ELMs respectively. Therefore, as individual ELMs are only considered in this article, the ELM period has been selected. Finally, a empirical criterion has been used. A simple visual inspection of the D_α waveform suggests the presence of higher frequency components in this signal for the case of type III ELMs.
Quantifying the latter aspect in a single value has not been a straightforward task. Different ways of representing the shape of the D_α emission with a single parameter have been tested (for example standard deviation of the Fourier spectrum, signal power of high frequency components, and crest factor). The best results in terms of success in the type I/type III classification process have been achieved with the so called Crest Measure (*CM*) that we have defined as a ratio between the Crest Factor (*CF*) and the number of Relevant Samples (*RS*) by means of:

$$CM = \frac{CF + Number\ of\ RS}{2}$$

CF is the crest factor that is defined as the ratio of the peak (crest) value to the Root Mean Square (*RMS*) value of a waveform.

$$CF = \frac{PEAK\ LEVEL}{RMS}$$

If the temporal evolution of the D_α signal during the ELM is represented by the samples $\{y_1, y_2,..., y_n\}$, then the *Peak Level* = max $\{y_1, y_2,..., y_n\}$ and the Root Mean Square is $RMS = \sqrt{\frac{y_1^2 + ... + y_n^2}{n}}$

The *Number of RS* is defined as the number of signal samples which exceed a *Threshold Level* (Fig. 2). This threshold has been empirically determined as

$$Threshold\ Level = (Peak\ Level + Final\ Level\ ) / 3$$

where the *Final Level* is the stationary level reached by the D_α signal.

3. CLASSIFICATION METHODS

After feature extraction, the classification methods must group the ELMs into two subsets: Type I and Type III. The creation of a classification system implies the use of training data and test data. Training data are needed to define the classes or categories that define the classification system. Test data are used to estimate the success rate of the classification system developed with the training data. JET signals have been analysed with the different techniques explained below.

3.1. SUPERVISED METHODS: SUPPORT VECTOR MACHINES

In supervised methods, both the number of classes (in the present case two classes, Type I and Type III) and the category for the training samples (each training data belongs to a well-known class: Type I or Type III) are known. Once the system is trained, it is ready to estimate the category of input data.

Support Vector Machines (SVM) has been used as supervised method for ELM classification. SVM is a universal constructive learning procedure based on the statistical learning theory [3]. The SVM maps input data into a high-dimensional space using a non-linear function. Once input data are mapped into the high-dimensional space, linear functions with constraints on complexity (i.e., hyper-planes) are used to discriminate the inputs, and a quadratic optimization problem must be solved to determine the parameters of these functions.

3.2. UNSUPERVISED METHODS

In unsupervised methods, only input data are given to a learning system and there is no notion of the category during learning. The outcomes of unsupervised methods are both the number of classes and the assignment of each training data. Two different techniques have been used with ELMs: K-means and hierarchical.

3.2.1. K-MEANS

K-*means [4] follows a simple and easy way to classify a given data set through a certain number of clusters (assume* k *clusters) estimated after the training process.*

The algorithm is composed of the following steps:

1. Place *K* points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the *K* centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

3.2.2. HIERARCHICAL

Given a set of *N* items to be clustered, and an $N \times N$ distance matrix, the basic steps of hierarchical clustering [5] are:

1. Start by assigning each item to a cluster, so that if you have *N* items, you now have *N* clusters, each containing just one item. Let the distances (similarities) between the clusters be the same as the distance (similarities) between the distances they contain.
2. Find the closets (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size *N*.

Of course there is no point in having all the *N* items grouped in a single cluster but, once the complete hierarchical tree has been obtained, *k* clusters can be estimated by cutting the *k*-1 longest links in the tree.

4. RESULTS

Each classification method presented in Section 3 (SVM, K-means and hierarchical) has been applied to 265 ELMs isolated from JET signals [6]. A total of 122 ELMs correspond to discharges used as training data (97 of type I and 25 of type III). Figure 3 summarizes the numerical ranges of each feature for this training set.

The test data consists of 143 ELMs. It is worthwhile to point out that similar results are obtained when using any of the classification methods, either supervised or unsupervised. In both cases the number of classes that provides the vast classification performance is two. Moreover, the success rates with different techniques shown in Table 1 allow concluding that feature selection adopted is quite robust. In particular, using K-means and hierarchical, the percentage of signals included in each cluster is the same.

With SVM we obtained one set of 238 ELMs of Type I and one set of 27 ELMs of type III. With K-means and hierarchical we obtained two set: 237 ELMs of Type I and 28 of Type III. Only one ELM belonging to the test data, and wrong classified by the unsupervised methods, is moved when using SVM from one class to another.

Figure 4 shows the results for the K-means classification that are very similar to the ones obtained with SVM and hierarchical. As it can be observed by inspection of the figure, ELMs of Type I have bigger drop than those of Type III but this characteristic is not good enough to absolutely differentiate between them. Moreover, the period is also not a deciding feature for the classification process. However classification depends strongly on the Crest Measure feature. In fact, the same success rate showed in Table 1 is obtained if just the Crest Measure is the only feature under consideration, *i.e.* no improvement in the classification is obtained by using the other parameters. One parameter, the Crest Measure, is enough to discriminate ELMs of Type I and Type III. On the other hand, period and drop cannot be used by themselves to obtain a good classification.

It should be emphasized in the analysis of the results that it is necessary to use the CM instead of the CF in the classification process. The CF is not good enough for the discrimination of the two types of ELMs. This fact denotes that the Number of RS is an important component in the CM feature.

CONCLUSIONS

In this paper we presented a method for the classification of ELMs based on individual ELM analysis. Although the physical basis to this approach is not established, the method seems to work on a restricted database of 300 ELMs. In addition, the selected dataset has been classified with very high success rates and a very low dimensionality (in fact in can be reduced to a single feature, the Crest Measure). The several methods used (supervised and unsupervised) group the same signals into the same sets which means that the features selected are robust enough to represent the ELM instability.

On the other hand, it should be noted that the database is not completely general and a comparison with a more extended database is needed in order to draw definitive conclusions on the approach proposed here. More difficult cases could require the development of additional analysis to take into account a wider range of conditions, for example power density and collision regimes.

ACKNOWLEDGEMENTS

This work, supported by the European Communities under the contract of Association between EURATOM/CIEMAT, was carried out within the framework of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission. The authors wish to thank Prof. Sebasti·n Dormido Bencomo (UNED) and Prof. Jes˙s Manuel de la Cruz (UCM) for their constructive comments and invaluable guidance. The authors wish to thank to E. de la Luna, A. Loarte, I. Nunes and E. R. Solano for their valuable comments.

REFERENCES

[1]. N. Duro, J. Vega, R. Dormido, G. Farias, S. Dormido-Canto, J. S·nchez, M. Santos, G. Pajares. Automated clustering procedure for TJ-II experimental signals. Fus. Eng. Des. **81** (2006) 1987-1991.

[2]. G. Saibene, L.D. Horton, R. Sartori, B. Balet, S. Clement, G.D. Conway et al. The influence of isotope mass, edge magnetic shear and input power on high density ELM and H modes in JET. Nuclear Fusion. **39**, 9 (1999) 1133-1156.

[3]. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, (1995).

[4]. MacQueen, B. Proc. of 5th Berkeley Symp. Math. Statistics and Probability, Berkeley, Univ. California Press, 1967, **1**, pp 281-297.

[5]. Johnson, S. C., Hierarchical Clustering Schemes, Psychometrika. 1967, **2**, pp 241-254.

[6]. http://users.jet.efda.org/pages/s1-task-force/index.html.

	Type I	**Type III**
K-means	98%	93%
Hierachical	98%	93%
SVM	98%	96%

Table 1. Rate of success in the classification.

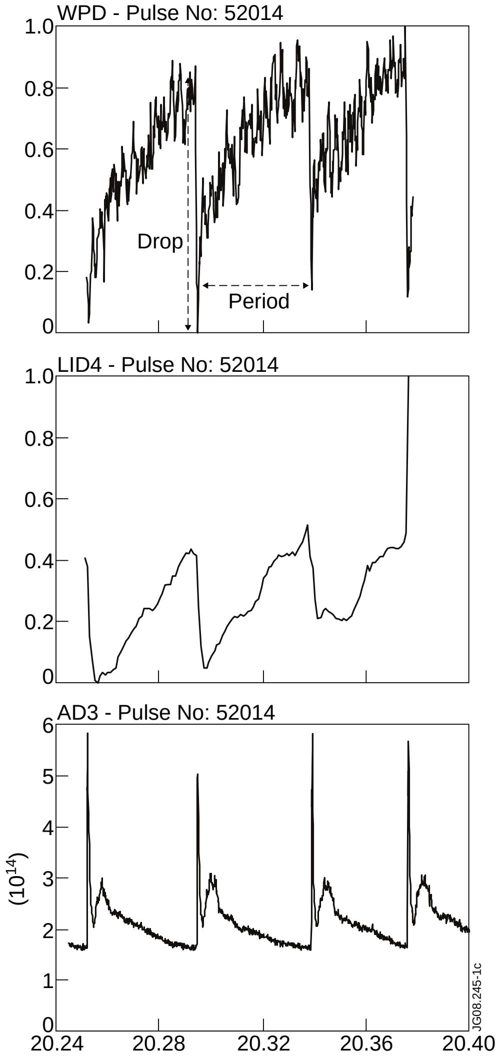

Figure 1: WPD, LID4 and AD34 signals for Pulse No:52014.

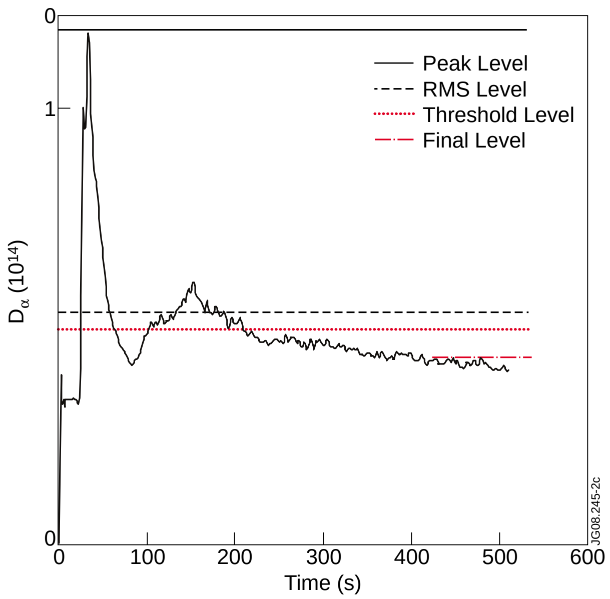

Figure 2: Crest Measure feature.

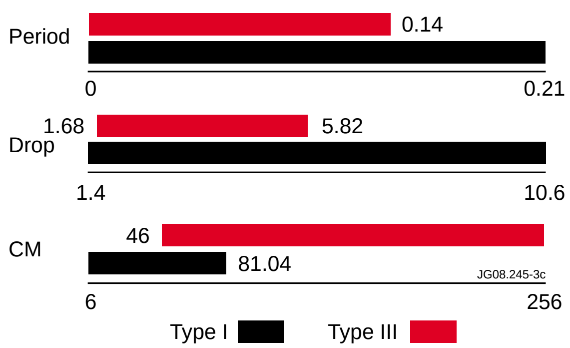

Figure 3: Variation range of features for the training set.

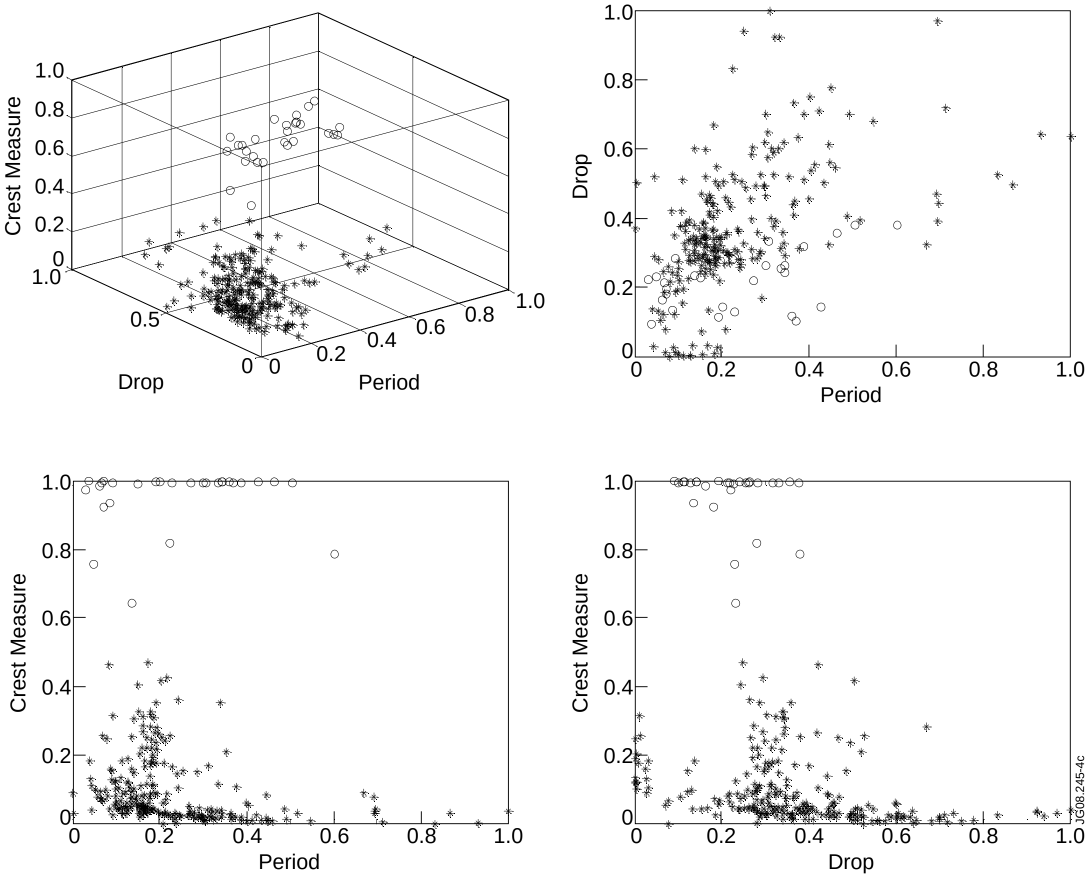

Figure 4: K-means results (Type I '' and Type III 'o').*