MACHINE LEARNING AT THE CTBTO. TESTING, AND EVALUATION OF THE FALSE EVENTS IDENTIFICATION (FEI) AND VERTICALLY INTEGRATED SEISMIC ASSOCIATION (VISA) PROJECTS

Ronan J. Le Bras[1], Stuart Russell[2], Nimar Arora[2], and Vera Miljanovic[1]

Comprehensive Nuclear-Test-Ban Treaty Organization[1] and the University of California at Berkeley[2]

ABSTRACT

Since 2009, an initiative to investigate the potential of machine learning methods to improve automatic data processing at the CTBTO and in particular the recall and accuracy of the automatic bulletins is starting to bear fruit beyond the stage of research and has entered the domain of development and testing with the goal of operational testing for one of the projects (FEI) by the end of 2011. The prospect for FEI is that the tool will comfort analysts in their decision-making process when they make decisions on whether a (mostly smaller) event is real or false, and it is thus an enhancement of the current analysis system. The VISA projects are more ambitious and aim at replacing key components of the processing system. The prototype of the first generation, which aims at replacing the current automatic association tool (GA), is being evaluated on the vDEC collaborative platform of the CTBTO. Results show much improved accuracy using VISA as compared to the SEL3 for the same recall value, or much-improved recall value using VISA as compared to the SEL3 for the same processing accuracy. A consequence is a significant decrease in either the number of false alarms or the number of missed events, depending on the setting of the processing parameters.

OBJECTIVES

The objective of this project is to evaluate the applications of Machine Learning techniques to the processing of waveform data at the IDC of the CTBTO in a quasi-operational environment. The ISS09 project initiated by the CTBTO in 2008 included a *Data Mining/Machine Learning* component, which was a new area of investigation for the organization (Russell et al., 2009). The following projects were tackled, classified according to the time scale at which they could become operational.

- The projects with operational short term goals included:
    - False Events Identification (FEI) using Support Vector Machine (SVM) methods (Mackey et al., 2009)
    - Hydroacoustic and Seismic phase identification (Tuma M. and Igel C., 2009; Schneider et al., 2010)

- The projects with operational medium-term and long-term goals included:
    - Vertically Integrated Seismic Analysis (NET-VISA and SIG-VISA) detection, association, and location (Arora et al., 2011a, 2011b)

The status of these different projects was presented in Le Bras et al. (2010). Two have reached a level of maturity sufficient to envision their integration into operations in the near future. One of them (FEI) is being tested on the development system of the International Data Centre. The other, NET-VISA, has been tested on the vDEC collaborative platform at the CTBTO (Vaidya et al., 2009) and has undergone improvements and testing over the last year. NET-VISA, which involves a paradigm change from the current operational framework, has reached the point that, after some adaptation to the operational environment and modifications to improve efficiency, the prototype is ready to be tested operationally within the next year.

RESEARCH ACCOMPLISHED

The various short-term and long-term projects tackled in the Machine Learning area during the last year have led to a number of publications illustrating the benefits that can be obtained from applying concepts in that field to the problem of processing of seismic and hydro-acoustic data at the IDC. In this paper, we report on two projects that approach actual implementation. They are the False Events Identification project, which attempts to identify whether the automatic event are likely or not to be valid LEB events, and the VISA project, which would replace the automatic association part of the current processing system.

False Events Identification

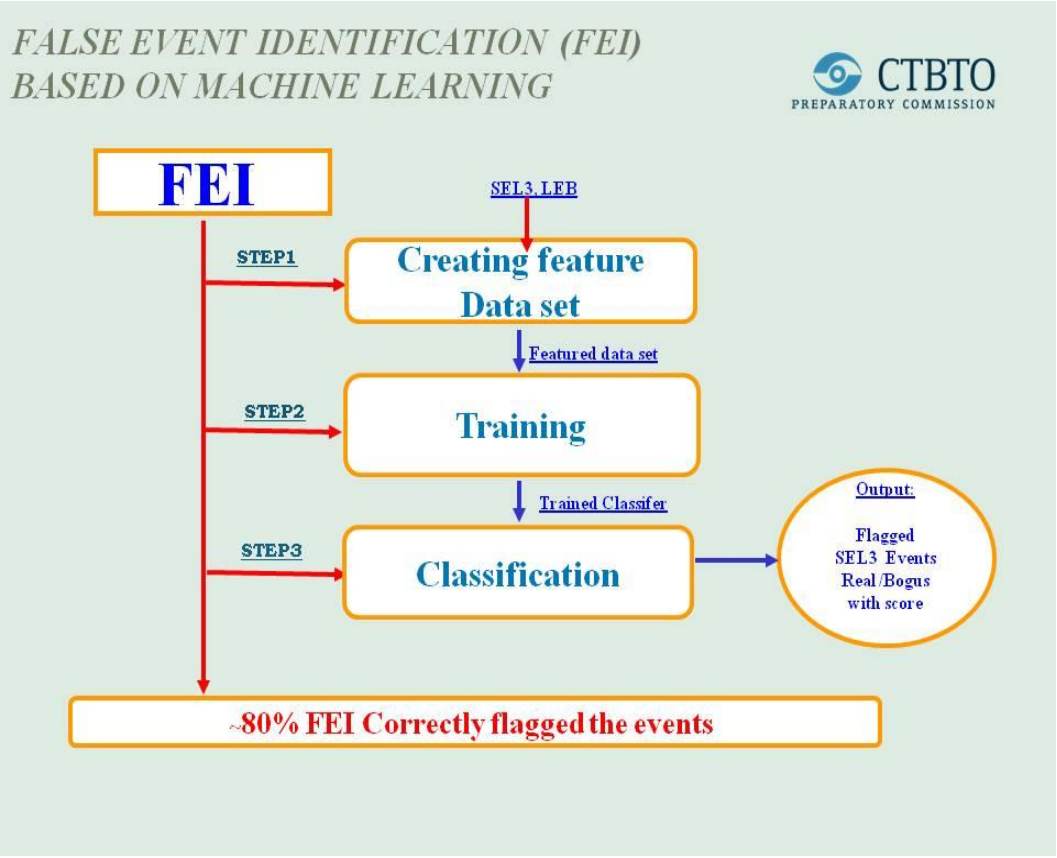
Background

The False Events Identification (FEI) program was written by Ariel Kleiner and Lester Mackey, of the University of California, Berkeley (Mackey et al., 2009). The program is written in Java.

FEI uses a Support Vector Machine (SVM) approach, with a very large feature set (e.g., Le Bras et al, 2010) to determine automatically whether or not each SEL3 event will be automatically discarded or retained by analysts. Using historical analyst-reviewed bulletins as "ground truth", FEI classifies each SEL3 event to be retained for further analyst review or to be discarded, with accompanying confidence score. The SVM algorithm provides a computationally efficient classifier whose accuracy is sufficient to enable a significant decrease in the false positive rate (false SEL3 events).

The approach is modularized into three parts:

- **Featurization** (feature selection). This first step simply extracts the fixed set of features from the input portion (SEL3 in this case) of the training data set.
- **Training**. This step provides the functionality for training classifiers. The classifiers are trained on SEL3 and LEB parametric data. The features extracted during the first step are used to predict the LEB event outcome (either real LEB event or a false alarm).
- **Classification**. Finally, in this step, which is to be used continuously in operations, the classifiers are used to predict whether or not a new event will be rejected by analysts.



The Java implementation of FEI is organized into three processing modes:

- create feat dset: creates a featurized dataset and writes the dataset to a file.
- train: trains a new classifier based on a featurized dataset, the result is saved to a file.
- classify: uses a trained classifier to generate predictions on new events. The predictions are written to a file in csv format or to a database table (this is the operational option and the one we are testing). Each event is given a score, and a label of either 1 (reject) or -1 (retain) based on the score.

A database table (EVENT_FEI_SCORE) has been designed to fit into the IDC schema and receive the FEI results.

Testing Method

In order to assess the variability of the results depending on the time period of the training data set, four FEI classifier files using varying time periods as training data were produced. After the classifier files were created, they were used to independently evaluate several periods between 2006 and 2010. May 1-7, 2010 was used for detailed evaluation using the two classifiers trained with data that did not include that time period. During the May 1-7, 2010 time period, there were 1050 events in the SEL3 database evaluated by the FEI classifier. The time periods used to train the classifiers are listed in Table 1. Several subsets of a larger time period were used for testing in order to evaluate the amount of data necessary to begin producing dependable results.

Start Date	End Date	No. of SEL3 / LEB Events used for Training the Classifier
February 24, 2006	March 05, 2006	1258 / 1161
April 1, 2010	April 8, 2010	1540 / 700
April 1, 2010	April 30, 2010	5514 / 2747
April 1, 2010	May 31, 2010	10,086 / 5174

Table 1: Training and Classifier Sets Created for FEI Testing

One method of evaluating the success of FEI processing was to track the event identifier (*evid*) from the automatically generated SEL3 database account and check for its presence in the human analyst reviewed LEB database account. With this method of evaluating the percentage of the time that FEI predicts the right answer (either correctly predicts false or correctly predicts actual LEB event), success rates generally exceeded 80 percent.

Results

When the data used for training and evaluations are closely spaced in time, FEI gives very good results: more than 80% of the FEI classifications are correct. Testing of the various classifier sets listed in Table 1 indicates that the larger the training set, the better the results (see Figure 1**: FEI results using two training data sets of different size on the same evaluation data set (May 1-7, 2010)**).

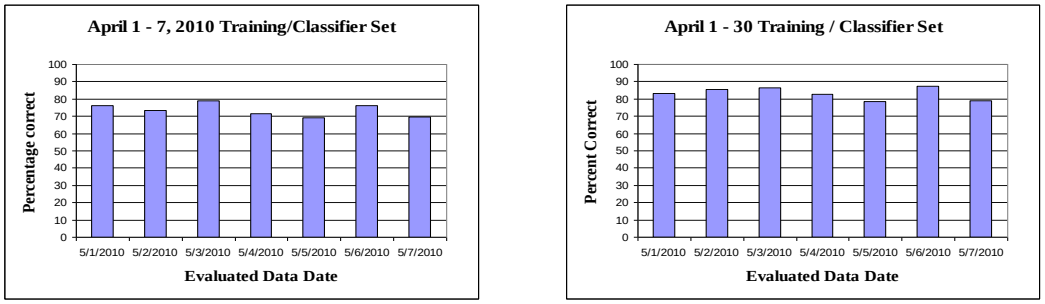


Figure 1: FEI results using two training data sets of different size on the same evaluation data set (May 1-7, 2010)

When the training and classification sets are from temporally separated time periods, a 25-30% degradation of performance was observed, regardless of the size of the training sets (see Figure 2). This degradation is attributable to variance in the composition of the network. When a new station is added to the network, the dynamics of event formation change. Likewise, if a station exists in the training set, but is not in the network of the evaluated data, a similar degradation is observed.

In Figure 2, the top chart shows that with a training set from 2010, and evaluation data from 2009, the performance is much better than the with a training set from 2006 on the same evaluation data. This is to be expected as the network has evolved between these dates. This has implications for the operational model to be used for this module. The training will have to be redone when network configuration changes, after sufficient data has been gathered with the new network configuration.

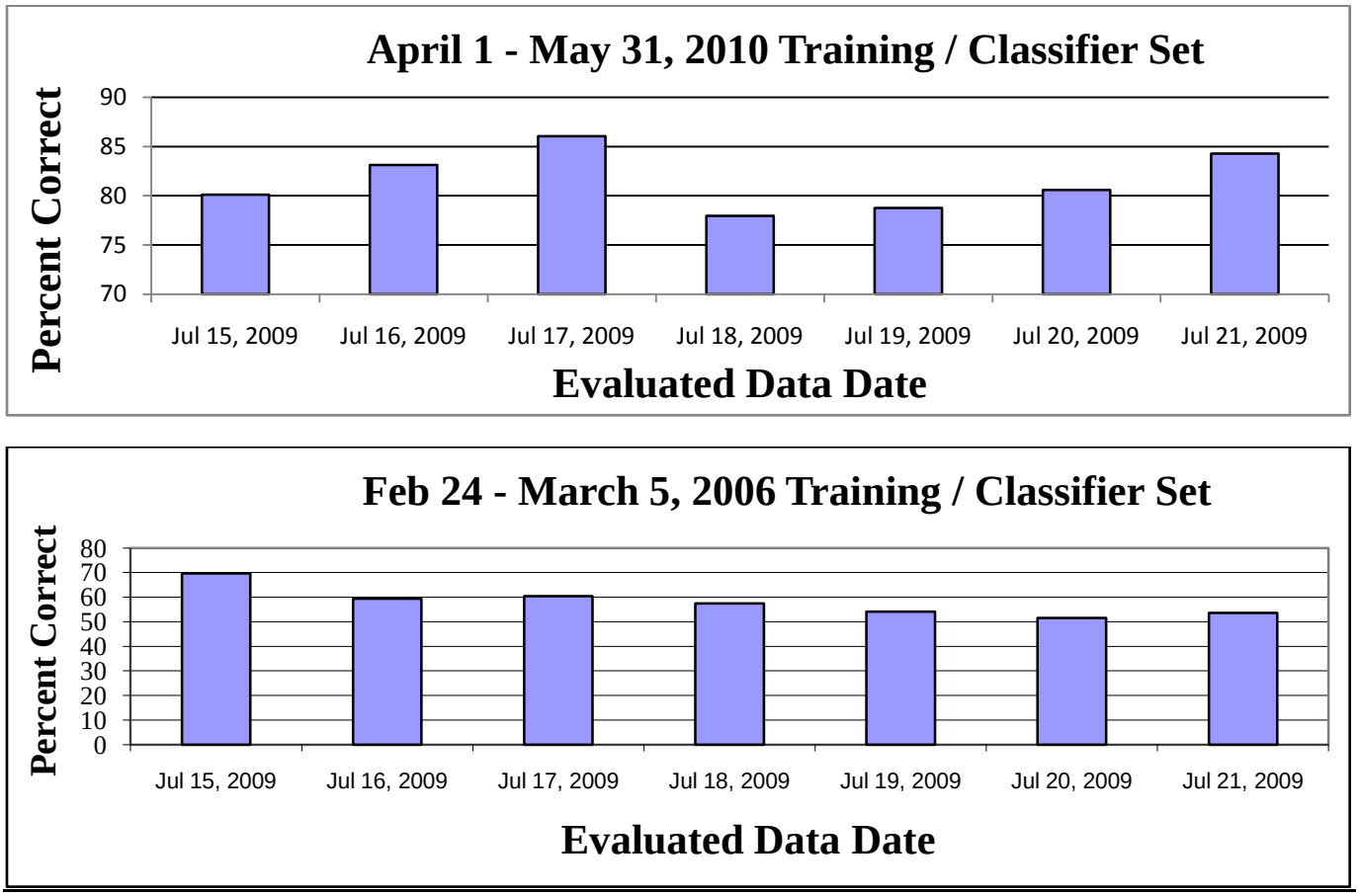


Figure 2: Using a trained classifier (with 2006 or 2010 data) against the same data. Note the better results when the training set is closer to the date on which the classifiers are applied. This is likely due to the changes in network configuration that occurred between 2006 and 2009, with 2009 being closer to the 2010 configuration on which the classifiers were trained.

Summary of testing

FEI does an excellent job at classifying/categorizing automatic events into either false events or events with a high probability of being real.

The results presented to analysts as the process currently stands should add confidence to their decisions and help identify obviously wrong associations and false events.

As it currently performs, FEI shows great promise. If additional optimization can be accomplished it will be a powerful tool to build on. One conclusion that stands out from our evaluation using different detector networks in the training and testing phases is that it is imperative that the operational model take into account changes in the

networks. When new stations are added, for instance, a new set of classifiers should be trained once sufficient automatic and analyst data has been accumulated to allow for the training to be performed.

NET-VISA improvements and testing

Evaluation on one week of data. March 22-29, 2009.

The NET-VISA project (Arora et al., 2011a, 2011b) is the first stage in the process of upgrading the automatic processing of seismic data from waveform processing to the production of automatic bulletins using Bayesian inference methods. In this first stage, the one-to-one replacement of the current automatic association process using as input the parametric detection data and ending with the production of an automatic bulletin is attempted. The project includes the production of a prototype and installation on the vDEC platform at the CTBTO. Offline tests have been performed on archive data and the prototype has been improved since the initial implementation of NET-VISA (Le Bras et al., 2010). The improvements include associating the detections marked as ***tx***, which are typically unassociated but are sometimes P (or other real) phases, and an improved model of noise amplitudes.

For the purpose of evaluation, the LEB bulletin was considered the ground truth and a comparison between the LEB bulletin and the automatic bulletin, SEL3 or NET-VISA, was made. Based on the matching, the precision (percentage of events in the automatic bulletin which are in the ground truth bulletin), recall (percentage of ground truth events which are in the automatic bulletin), and average error (distance between a ground truth event and the matching automatic bulletin event) were obtained. Table 1 shows the recall and average error of SEL3 and NET-VISA, while Figure 3 shows the precision-recall curve for the latest implementation of NET-VISA as well as the earlier implementation. Because the LEB "ground truth" is derived by human analysts from SEL3, the comparison between VISA and SEL3 is likely to be biased in favor of SEL3.

Table 1. Breakup of SEL3 and NET-VISA performance by m_b.

m_b range	Total number of events	SEL3			NET-VISA		
		Recall (%)	Error (km)	Standard Deviation (km)	Recall (%)	Error (km)	Standard Deviation (km)
0-2	74	64.9	101	107	86.5	101	100
2-3	36	50.0	186	167	77.8	159	129
3-4	558	66.5	104	117	86.4	115	113
>4	164	86.6	70	112	93.3	78	108

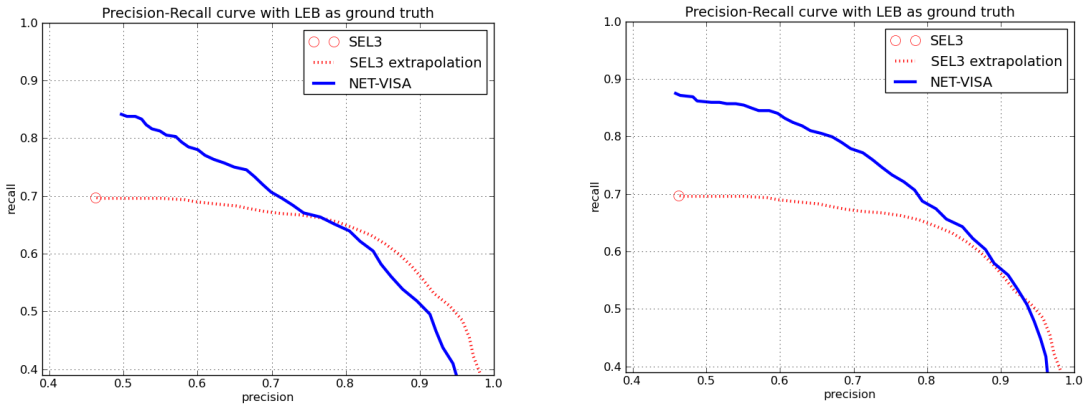


Figure 3: Precision-Recall curves for an early implementation of NET-VISA (left curve), as presented in Le Bras et al., 2010, and the latest implementation of NET-VISA (right curve). Note the better recall results at higher precision with the latest implementation.

Comparison with non-IDC bulletins as ground truth

All the previous results are based on the assumption that the LEB bulletin is the ground truth, which is not completely correct; while the bulletin produced by the IDC analysts is of high quality considering the sparseness of the IMS network and the limited amount of time available to produce it, it is not exactly the ground truth, especially for smaller events. For the one week period analyzed, the following observations can be made:

- In the continental United States of 33 events reported by NEIC:
  - LEB got 4 correct out of 4 predicted events
  - NET-VISA got 7 correct out of 35 predicted events
- In Japan out of 1565 events reported by JMA:
  - LEB got 29 correct out of 29 predictions
  - NET-VISA got 33 correct out of 52 predictions
- In Europe out of 101 events reported by PRU
  - LEB got 5 correct out of 10 predictions
  - NET-VISA got 11 correct out of 43 predictions
- In Central Asia out of 101 events reported by NNC
  - LEB got 35 correct out of 74 predictions
  - NET-VISA got 50 correct out of 166 predictions

These results are quite interesting since they suggest that the replacement of the current automatic association program by NET-VISA would lower the missed event rate significantly for smaller events and it is interesting to speculate about the effect on the LEB bulletin that this would eventually have, since the analysts are likely to be influenced by the bulletin used as input to their analysis. It would be a very costly experiment to run the two automatic methods side by side for a long period of time, but it would be quite feasible to do this for a period of a few days, perhaps up to a week, with two equally seasoned analysts involved in the processing.

The DPRK event of 25 May 2009.

The prior events distribution model used in NET-VISA includes two parts. One is based on the observed seismicity and will tend to place newly formed events in areas of previous seismicity. The other part is a spatially uniform distribution, i.e., it allows for an event to occur at any place on the surface of the Earth with equal probability. In

order to verify that nuclear explosions will be obtained correctly by the process, it was tested on the DPRK event of 25 May 2009 (for this experiment NET-VISA was trained on a one-year dataset from April 1, 2008 to April 1, 2009), and it was verified that the event was obtained correctly. Figure 4 shows the relative locations of the SEL3, LEB, NEIC, and NET-VISA events.

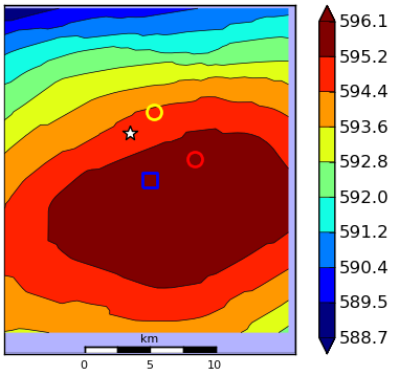


Figure 4: Relative locations of the SEL3 (red), LEB (yellow), NEIC (white star), and NET-VISA (blue square) events. The events are plotted in the event location density background for NET-VISA.

The DPRK event was detected at 39 stations by SEL3, while NET-VISA detected it at 53 stations using the same automatic detections, and LEB also detected it at 53 stations (50 of which were common with the NET-VISA detections). However, LEB was able to detect the event at an additional 8 stations using detections that the analysts added by hand.

CONCLUSIONS AND RECOMMENDATIONS

Two programs resulting from the machine learning efforts at the CTBTO are in the process of being evaluated for their possible installation in IDC operations. The FEI program is the closest to operational implementation and has been installed on the development system of the CTBT after successful evaluation on the vDEC platform. Testing has been performed and has resulted in a better understanding of what the operational model should be for this component of the system. NET-VISA reduces detection failures by more than a factor of 2 compared with SEL3, and this is a significant achievement in itself. Most of the events on which the evaluation of NET-VISA has been done are earthquakes, since these constitute the overwhelming majority of events seen by the IMS network. It is not surprising that the performance on natural events is improved, since the NET-VISA Bayesian method includes prior statistics learned from the archive data. It was shown however that the event of the 25th May 2009, the second announced nuclear test from the Democratic People's Republic of Korea, was obtained by NET-VISA. This is a verification that the complete prior model, which includes a uniform spatial distribution in addition to the seismicity-dominated prior, is adequate to detect events which do not occur in areas of previous seismicity. NET-VISA is currently being tested in the CTBTO vDEC environment for possible deployment in operations. It is necessary that more test cases be evaluated on the vDEC platform and that seasoned analysts have access to the results in order to evaluate them from their point of view. The next step in terms of algorithmic development is to develop the SIG-VISA prototype with an extension of the generative model down to waveform level, and include the step of signal detection within the generative model.

ACKNOWLEDGEMENTS

We thank Dr. Lassina Zerbo, IDC Director, for allowing us to publish this research and for his support of the machine learning efforts at the IDC, Ronald "Chip" Brogan and Misrak Fisseha for their analyst expertise and efforts in evaluating the FEI software delivered at the IDC.

REFERENCES

Arora, Nimar S., Stuart J. Russell, Paul Kidwell, and Erik Sudderth, "Global seismic monitoring: A Bayesian approach.", In *Proc. AAAI-11*, San Francisco, 2011.

Arora, Nimar S., Stuart J. Russell, Paul Kidwell, and Erik Sudderth, "Global seismic monitoring as probabilistic inference.", In *Advances in Neural Information Processing Systems 23*, MIT Press, 2011.

Arora, Nimar S.., Russell, S., Kidwell, P., and Sudderth, E., "NET-VISA model and inference improvements", CTBT:S&T 2011 Conference, Vienna, Austria, June 8-10, 2011.

Mackey, L.; Kleiner, A.; Jordan, M. I., 2009, Improved Automated Seismic Event Extraction Using Machine Learning**,** American Geophysical Union, Fall Meeting 2009, abstract #S31B-1714.

Tuma, M. and Igel, C., 2009, Kernel-based machine learning techniques for hydroacoustic signal classification, CTBTO ISS09 Conference.

Le Bras, Ronan J., Sheila Vaidya, Jeffrey Schneider, Stuart Russell, and Nimar Arora, "Status of the Machine Learning Efforts at the International Data Centre of the CTBTO." In *Proc. Monitoring Research Review (MRR 2010)*, Orlando, Florida, 2010.

Russell, Stuart, Sheila Vaidya, and Ronan Le Bras, "Machine Learning for Comprehensive Nuclear-Test-Ban Treaty Monitoring.", *CTBTO Spectrum*, 14, 32-35, 2010.

Vaidya, S., Robert Engdahl, Ronan Le Bras, Karl Koch, and Ola Dahlman, 2009, Strategic Initiative in Support of CTBT Data Processing: vDEC (virtual Data Exploitation Centre), CTBTO ISS09 Conference (http://www.ctbto.org/fileadmin/user_upload/ISS_2009/Poster).

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the CTBTO Preparatory Commission.