Big Data Analysis Techniques for Cyber-Threat Detection in Critical Infrastructures

William Hurst, Madjid Merabti, Paul Fergus
PROTECT: Research Centre for Critical Infrastructure Computer Technology and Protection
School of Computing and Mathematical Sciences,
Liverpool John Moores University,
Byrom Street,
Liverpool, L3 3AF, UK
w.hurst@2009.ljmu.ac.uk, {m.merabti, p.fergus}@ljmu.ac.uk

***Abstract* — The research presented in this paper offers a way of supporting the security currently in place in critical infrastructures by using behavioural observation and big data analysis techniques to add to the Defence in Depth (DiD). As this work demonstrates, applying behavioural observation to critical infrastructure protection has effective results. Our design for Behavioural Observation for Critical Infrastructure Security Support (BOCISS) processes simulated critical infrastructure data to detect anomalies which constitute threats to the system. This is achieved using feature extraction and data classification. The data is provided by the development of a nuclear power plant simulation using Siemens Tecnomatix Plant Simulator and the programming language SimTalk. Using this simulation, extensive realistic data sets are constructed and collected, when the system is functioning as normal and during a cyber-attack scenario. The big data analysis techniques, classification results and an assessment of the outcomes is presented.**

Index Terms— Critical Infrastructure, Big Data, Behavioural Observation, Simulation, Data Classification

1. Introduction

Critical infrastructures include sectors such as energy resources, finance, food and water distribution, health, manufacturing and e-government services [1]. Their service provision is often dispersed over large geographic areas [2]. In recent years, critical infrastructures have become increasingly dependent on ICT to facilitate communication. Consequently, this makes these systems more vulnerable and increases the threat of cyber-attack from different sources [3]. Our research, to date, involves the use of Behavioural Observation for Critical Infrastructure Security Support (BOCISS) [4]. Our observer system monitors an infrastructure's behaviour and detects abnormalities, which are the result of a cyber-attack taking place.

The system uses mathematical classification techniques to evaluate datasets and detect changes in behavioural patterns. By observing subtle changes in system behaviours, an additional level of support for critical infrastructure security is provided. The results achieved during the data classification process are encouraging. The data used in the evaluation of our system is produced by a simulation of a critical infrastructure. BOCISS analyses the data produced and identifies the behavioural patterns.

In this paper, an overview of the BOCISS system design, big data classification techniques and the data processing methodology is presented. The results are presented in two stages. Firstly, a smaller feature set and data sample provides an insight into the classification performance. The second part of the approach involves a larger dataset with an increase in the number of features used. A discussion on the sets of results is also provided.

2. Data Construction

BOCISS requires realistic critical infrastructure data. This is provided by the development of a nuclear power plant simulation using the Siemens Tecnomatix Plant Simulator and the programming language SimTalk. Using this simulation, realistic data is constructed and collected, when both functioning as normal and during a cyber-attack scenario. Figure 1 displays an overview of the whole system which is known as a pressurised water nuclear reactor [5].

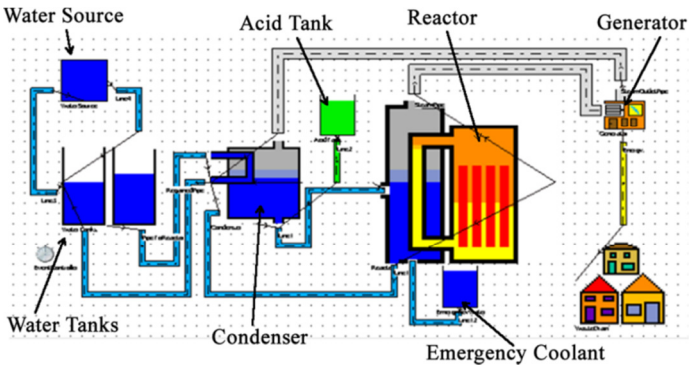

Fig 1. Infrastructure Simulation

Each of the mechanisms has a graphical icon to represent its function more clearly. They can also be expanded to detail their interconnectivity and the various components, which allow the system to operate. Each of the mechanisms are explained.

- The Water Source: The production of water is supplied by three sources including two infinite sources, representing a lake or ocean, and one water tower. The water requires filtering. It is, therefore, supplied to one large pipe, which is then sifted for impurities before being pumped into the Water Tanks.
- The Two Water Tanks: Water produced by the water source is collected in two tanks. This effectively acts as a buffer and controls the water flow in the system. Two pumps are used to send water from the tanks to the Condenser.
- The Condenser: One of the most complex mechanisms in the simulation is the condenser which consists of an interaction between two system loops. In the condenser, steam is cooled and converted to water and the water is subsequently sent to the reactor to be heated.
- The Reactor: The intake of water is combined with heat from a nuclear reaction to produce steam in the reactor mechanism. The steam is sent to the generator mechanism via two steam pipes.
- The Generator: Steam sent from the reactor rotates a turbine. Each unit of steam turns the turbine once and energy is produced. Excess steam is directed via a network of pipes to the condenser system for cooling.
- Acid Tank and Emergency Coolant: In case of system failure, two storage tanks, one containing Boronic Acid and one containing emergency coolant, are in place. The emergency coolant is required in the case of a failure in the provision of water to the reactor. The Acid is needed for emergency situations such as core overload as a result of cascading system failure.

Each component in the simulation has a corresponding observer, which extracts physical information about behaviour and constructs the data set required for the BOCISS evaluation. In order to have successful data classification, both normal behaviour data and attack data is needed and constructed in our simulation. The normal data set was constructed by running the simulation for a period of two simulated days with active sampling conducted at 4Hz (which is every 0.25 of a second). Therefore, the dataset generated consists of 732,000 records of data for each component.

3. Data Analysis Approach

The constructed data set is processed by our system which uses various mechanisms and data stores. (A full detailed account of the system can be found in [4]). The system connects directly to the network and registers itself and begins data collection in blocks.

A. *System Design*

Extracted network data is converted in the data manager and, depending on the mode of operation, is directed to the data store or for feature extraction. Features, initially sent to a temporary data store are used to create feature vectors and train classifiers to identify system behaviours. A system control governs the operations and interprets the classification results for the UI. The system design is displayed in Figure 2.

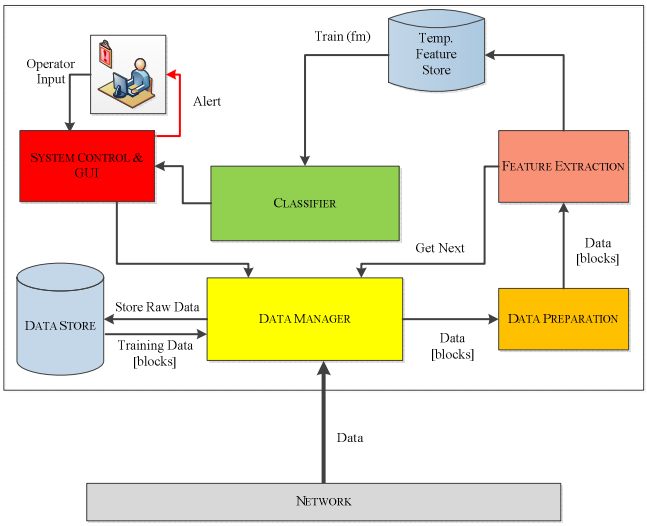

Fig 2. BOCISS System Overview

The use of a data manager enables BOCISS to act as a plug-in service. The data manager uses a data acquisition application (DAQ) and interprets the protocol-formatted data extracted from the network. A data acquisition application is constructed from a hardware component complete with a piece of software, which extracts data from a source. Two examples of protocol data formats include DNP3 and Modbus. Both are used if the critical infrastructure uses a SCADA system [6], [7]. The data manager converts it to raw data using the data acquisition application and sends it either to a database, which is able to store both normal and threat behaviour separately, or to a feature extraction process.

Data I/O ⟶ DAQ Services ⟶ Raw Normal/Abnormal Data

B. *Feature Extraction*

Features are aspects of the data, which allow for a representation of overall system behaviour [8]. In the training mode, features are extracted to form feature vectors for both normal and abnormal behaviour. The feature vectors are then stored in a temporary feature store until the data processing is complete. Once all the required data has been processed, a signal is sent to inform the temporary feature store to transfer its contents to the data classifiers. The features selected are unique for each critical infrastructure but they could include, for example, aspects such as: overall water volumes; steam output; energy creation; water tank levels or speed of water flow. They are constructed by cataloguing the data into designated representations of the dataset.

C. *Classifiers*

Feature vectors are constructed from the extracted features. The evaluation process uses supervised learning, by employing the feature vectors. The approach involves specific data classification techniques including: Uncorrelated Normal Density based Classifier (UDC), Quadratic Discriminant

Classifier (QDC), Linear Discriminant Classifier (LDC), Decision Tree (TREEC), and Parzen Classifier (PARZENC).

Linear Discriminant Classifier (LDC), is a technique which works by sorting or dividing data into groups based on characteristics to create a classification [9]. A discriminant function is obtained by monotonic transformation of posterior probabilities [10]. In other words, it performs an ordered transformation of unknown quantities, which are separated by a linear vector. Quadratic Discriminant Classifier (QDC) works in a similar way to LDC by dividing the data into groups based on given characteristics. However, by using QDC the data is divided using a quadratic surface rather than a one-dimensional one. QDC makes no assumptions that covariance are alike. In other words, it assumes that the changing of two random variables will not be the same [11].

Uncorrelated Normal Density based Classifier (UDC) also operates comparably to the QDC classifier but computation of a quadratic classifier, between the classes in the dataset, is done by assuming normal densities with uncorrelated features. Quadratic Bayes takes decisions by assuming different normal distribution of data [12]. LDC, QDC and UDC are density based classifiers. Decision Tree (TREEC) is a classifier which uses decision rules to divide the classes of data [10]. It operates by using criterion functions (the sum of squared errors), stopping rules (criteria for appropriate number of splits in a decision tree) or pruning techniques (the removal of unwanted tree sections). Using decision tree is a particularly ideal choice of classifier because it is well-known as one of the most effective supervised classification techniques [11]. Parzen Classifier (PARZENC) functions by including aspects of the training data when the classifier is built up. It is a non-linear classifier and it has the benefit that its parameters can be user supplied or optimised [10], [11].

4. Initial Classification

Using the above classifiers, the goal of the initial classification process is to demonstrate the techniques for abnormal behaviour detection using five data classifiers. The features selected for input into the data classification algorithms are based on an evaluation of which extracted characteristics provide a true representation of our simulation's behaviour. The features used, therefore, include aspects such as regular occurrences in system behaviour, and traits from individual components.

The features selected represent characteristics of system behaviour [13]. The features include, for example, 128 mechanism component features and 36 system component features. The system components are comprised of pipes or cables, which link the mechanisms together. The features are constructed by taking the maximum, minimum, mean and median values every hour from the data which is sampled at 4Hz (4 times every second) for a 24 hour simulation.

Each of the mechanism components provide a value produced by sampling the level of water, steam or energy passing through. This is also the case for the system components. 32 mechanism components provide 4 features each to form 128 features. The nine system components also provide 4 features each to produce 36 features. Using the features extracted from the two datasets for normal and attack behaviour records are created. These initial records of data are used for testing the classifiers' ability to identify normal behaviour and, subsequently, recognise when normal behaviour is not occurring. In total, 12 feature vectors were used to train the classifiers consisting of 6 for normal behaviour and 6 for abnormal behaviour.

Minimum, maximum, mean and median values were selected to form our initial feature vector records because each provides an ideal representation of the system behaviour. For example, when observing the minimum and maximum levels of water in a pipe or water tank, the constraints of normal behaviour can be specified. If the levels recorded are lower or higher than the expected minimum or maximum values then the system is not behaving as it should. In the same way, observing the mean levels of water steam or energy allows us to identify the normal behaviour constraints of selected critical components. Table 1 presents a sample of the initial record set, which consists of an evenly divided dataset randomly divided using MATLAB into a 50% training set, with the rest of the 50% assigned to a test set.

Table 1 Initial Data Set Sample

W2WTP	median	Max	min	W3WS1	median	Max	min	W4WP1	median
0	0	0	0	1	1	1	1	11	11
0	0	0	0	1	1	1	1	10.6	11
0	0	0	0	1	1	1	1	11	11
0	0	0	0	1	1	1	1	10.2	10
0	0	0	0	1	1	1	1	10.8	11
0	0	0	0	1	1	1	1	10.8	11
0	0	0	0	1	1	1	1	0	0
0	0	0	0	1	1	1	1	0	0
0	0	0	0	1	1	1	1	0	0
0	0	0	0	1	1	1	1	0	0
14	14	14	14	1	1	1	1	11	11
14	14	14	14	1	1	1	1	11	11

The first six vectors, in blue, represent normal data, whereas, the red values represent abnormal data. Using the above data sample, the performance of each classifier is evaluated to assess the classification accuracy. In the following subsection, an evaluation of the results is presented.

A. Initial Dataset Evaluation

In order to obtain a more accurate assessment of which of the classifiers is most successful and consistent, the experiments were conducted 30 times. Statisticians identify that experiments conducted 30 times provide an adequate realistic average [14]. An overall evaluation of the classification algorithms is presented in Table 2 which displays the results of classification success, sensitivity and specificity with the mean value taken for 30 experiments.

Table 2 Average Classifier Performance for Initial Dataset

Classifier	Classification Success	Sensitivity	Specificity
LDC	62.78%	1	0.21
UDC	88.90%	1	0.82
QDC	50%	1	0
PARZENC	50%	1	0
TREEC	90.01%	1	0.8

The results for sensitivities, which is the identification of normal behaviours, are high. The specificities, which is identification of abnormal behaviour, is mixed. Several of the classifiers are prone to generating false positive results, meaning abnormal behaviour values are grouped with normal behaviours. It is clear from the results that the classifiers are able to identify normal behaviour with high success, as displayed by the sensitivity results. However, all of the classifiers are prone to errors when detecting abnormal behaviours, with QDC and Parzenc failing to identify a single abnormal behaviour.

B. Initial Results Visualisation

In this subsection, a visualisation of the results is presented. Each diagram represents a sample of the outcomes and provides a visual demonstration of how the classifiers function. In each figure, the division of data into two groups for normal and abnormal behaviour is displayed. Firstly, Figure 5 shows the mapping of two classes on a scatter plot in a 2-D feature space for the Parzenc analysis.

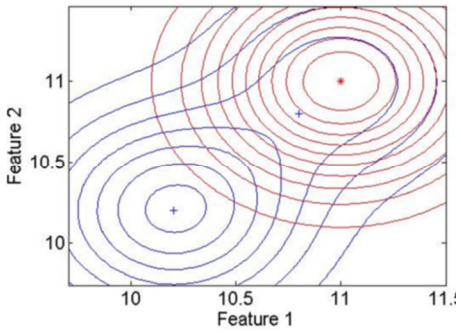

Fig. 5 Parzenc Plot 2 Features

Feature 1, on the x-axis, refers to one of the dominant features and Feature 2, on the y-axis, refers to one of the lesser dominant features from the dataset. Two features were used in each visual representation to demonstrate how the classifiers function. The ellipses, displayed, refer to likelihood contours, where the points inside the ellipse are most likely to belong to that grouping. The blue ellipses consist of data that comes from the normal behaviour dataset and the red ones referring to threat behaviour data. Threat behaviour can be identified as a result of one grouping clearly standing out from the other. As the graph displays, Parzenc struggled to cluster the data into its correct grouping as the ellipses for normal behaviour values contour the red abnormal behaviour values.

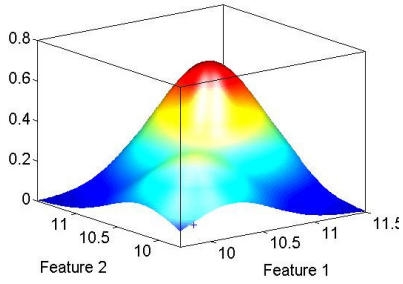

Fig. 6 Parzenc Plot 2 Features 3D

Figure 6, again shows the Parzenc results by mapping them in 3-D, where the ellipses are displayed as spikes or curves in three dimension. Ideally two clear spikes should be visible to demonstrate two distinct data groupings.

C. Initial Summary of Results

The initial results of the classification conducted using a small dataset, support our findings that data classification can be used to detect abnormal behaviour in critical infrastructures. Whilst the results show that anomalous behaviour can be identified with some success using our chosen classifiers, our initial dataset impacted the results. The ability to classify abnormal behaviour, was hampered by the fact that our dataset was too small to allow the classifiers to train themselves to a substantial level. The results will be expanded upon in the following section. We purposefully selected an abnormal dataset, which had a mix of substantial deviations from the normal behaviour as well as similar values. The results show that the classifiers where able to achieve a high success rate when identifying normal behaviours.

5. BIG DATA ANALYSIS TECHNIQUES

Building on the results from the initial data set evaluation, additional features are taken into consideration when classifying the larger dataset. Using a more substantial dataset, in this section we present a more conclusive evaluation of the selected classifiers. The dataset used in this section, has a large number of more subtle anomalies in the behavioural data in contrast with the initial dataset.

A. Classification Evaluation

A comparison of the classification success for each of the classifiers is presented in Table 4 below. Overall, the algorithms were able to accurately classify 96.653% of the dataset on average between them.

Table 4 Classification Results

Classifier	Classification Success %	Sensitivity	Specificity
LDC	93.64	0.99957	0.874
UDC	99.759	1	0.995
QDC	89.868	1	0.798
PARZENC	100	1	1
TREEC	100	1	1

The results presented are a significant improvement on the initial evaluation. LDC, QDC and UDC have mixed results; however, each also displays a significant ability to accurately classify behaviour. As previously, the classifiers are able to identify normal behaviour with high success with nearly all the errors occurring for the misclassification of abnormal behaviour. In the following subsection, we present a visualisation of the results, as well as, a discussion and a justification of the outcomes.

B. Main Results Visualisation

As with the initial classification, the visualised results presented represent a sample of the classification outcomes.

Figure 7 displays a scatter plot of two classes in a 2-D feature space for the LDC analysis. For the purposes of visualising the results, two of the features from the set are plotted once more.

The graph displays normal behaviour, represented by the blue cluster, and abnormal behaviour visible in the red cluster. The linear line generated by the LDC analysis displays the division between the two sets of data. Figure 7 displays one of the 30 experiments for LDC, which, on that occasion, obtained 100 % success.

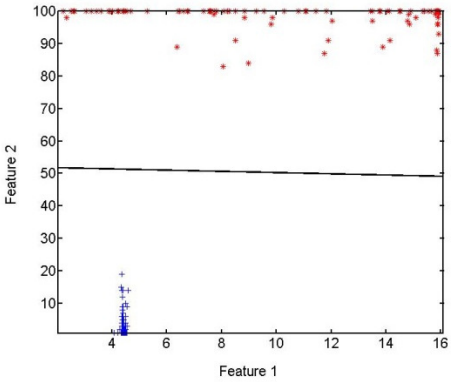

Fig. 7 LDC Analysis Graph 100%

Figure 8 displays the same scatter plot, however, each of the classifiers' approach for dividing the data are visible. For this particular experiment, all the nine classifier are able to divide the data into two distinct clusters accurately. The diagram displays a clear visualisation of the methodology for each classifier when dividing the data. As in the initial evaluation, a visualisation of the Parzenc classification is displayed in Figure 8.

However, in this case, Parzenc classification achieved higher results. As before, the blue ellipses consist of data that comes from the normal behaviour dataset and the red ones referring to threat behaviour data.

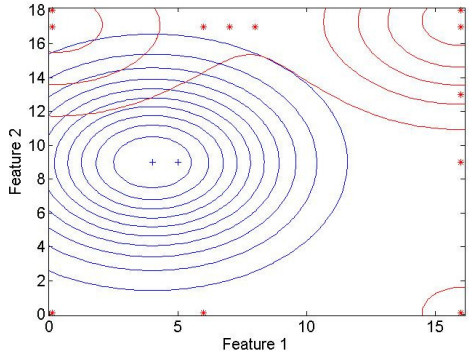

Fig. 8 Parzenc Evaluation 2 Features

The ellipses for normal and abnormal behaviour values more accurately contour the correct data clusters than previously. This is again displayed in 3D, in Figure 9, where two distinct peaks created by the data groupings are visible. As with the initial evaluation, the results obtained support the findings that data classification can be used to detect abnormal behaviour in critical infrastructures. In light of this, in the following subsection, we present an assessment of the results obtained using the main dataset.

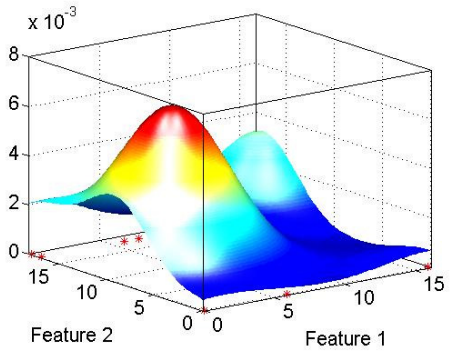

Fig. 9 Parzenc Evaluation 2 Features 3D

The classifiers are able to identify normal behaviour easily, with errors occurring for the classification of abnormal behaviour in the wrong cluster. However, the amount of misclassified data is relatively low. In the following subsection, a discussion on the results obtained in both of the evaluation stages is presented, along with a comparison of both.

6. DISCUSSION

The success of the classifiers is a result of various key stages including, noise reduction and principal component analysis prior to the classifiers being applied. In this section, we present a discussion and justification of the results obtained during the evaluation process.

A. Results Comparison

One of the main observations is that the increased amount of data improved the results. This is reflected in the comparison between the initial results and the main supervised machine learning results. Figure 10 displays a visual comparison of the mean classification success of each of the classifiers combined between the first and second evaluation.

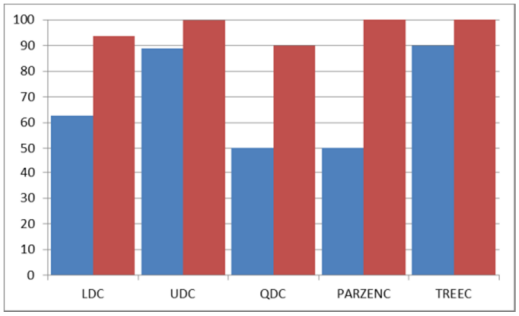

Fig. 10 Results Comparison

The blue bars represent the initial evaluation, whereas the red bars signify the main evaluation results. In the initial evaluation, the classifiers were able to classify 68.34% of the data accurately. This is lower than what would be ideal, for critical infrastructure security. In the case of critical infrastructures, it is important to achieve a high success rate.

Subsequently, in the second evaluation, the classifiers were able to achieve a much higher mean rate of 96.653% correctly classified data.

Providing the classifiers with more data, allows the algorithms to be trained to an effective level. There is a notable improvement in the overall classification results. The results display a remarkable improvement, in particular for LDC, QDC, UDC and Parzenc between the initial and main evaluation process.

B. Results Discussion

The improvement in the classification results is impacted by two key differences in the approach. Firstly, an advanced feature extraction process enhanced the outcomes. Collecting more features, which provided a more comprehensive representation of system behaviours, allowed the data to be filtered before being processed by the classifiers. Secondly, having a larger dataset to train the classifiers produces results that are more accurate. As reflected by LDC, UDC and QDC in particular, which have been significantly improved upon.

Initially, LDC, QDC and UDC performed less effectively. However, both were able to classify a significant proportion of the data accurately. The evaluation presented in this chapter demonstrates how normal behaviour can be identified in a system using data classification techniques. The results achieved were high and successful. Our evaluation is affected by; firstly, the quality of data used which had an impact on the results. Despite creating a simulation which replicates a critical infrastructure, the quality of data produced is inferior to a real-world nuclear power plant. The dataset generated is intended to demonstrate the ability of the classifiers to classify data into its correct groupings and identify when any given system behaviour is not as it should be.

Secondly the high results were achieved through an efficient pre-processing of the data which removed noise and selected the most effective features for training the classifiers. It was also apparent that we 'over attacked' the system creating a mixture of large and more subtle anomalies in the data. The principal component analysis stage selected several of the features with large data anomalies for training the classifiers, in addition to the features with more subtle anomalies.

Finally, the advantage of using supervised machine learning had an impact on the results we achieved. As previously discussed, the approach involved giving the classification algorithms the 'right answer' to enable them to operate self-sufficiently. By using this method, we are able to train the classifiers using features which are known to be effective for achieving high results.

7. CONCLUSION AND FUTURE WORK

Critical infrastructures are growing in size and importance every year as the population grows and puts increasing demand on the unseen services provided. Protecting these infrastructures is clearly a key issue. Improved support, as well as helping with cost efficiency as billions are spent on cyber security, has benefits for the well-being of people and helps with the evolution and improvement of security. As threats increase it becomes clear that security may lie away from conventional computer security techniques and an original approach to critical infrastructure protection is required.

The research presented in this paper presents the effectiveness of BOCISS. The classification techniques used present a demonstration of how our system is able to support security by applying big data analysis techniques to identifying anomalous behaviour caused by cyber-attacks taking place.

Future work will involve the adaptation of BOCISS to identify specific attacks, in addition to its current behavioural observation services. This will be done by recognising known behaviour changes, and what is causing them by drawing from system information stored in a database. This approach differs from signature-based detection as it looks at physical changes in component behaviour and uses them to identify attacks, which are known to cause those changes. Traditional signature-based detection identifies known data signatures, such as globally known viruses.

REFERENCES

[1] M. Merabti, M. Kennedy, and W. Hurst, "Critical infrastructure protection: A 21st century challenge," in 2011 International Conference on Communications and Information Technology (ICCIT), 2011, pp. 1–6.

[2] Á. MacDermott, Q. Shi, M. Merabti, and K. Kifiyat, "Considering an elastic scaling model for cloud security," in International Conference for Internet Technology and Secured Transactions (ICITST), 2013.

[3] N. Nicholson, "SCADA Security in the light of Cyber-Warfare," Elsevier Comput. Secur. J., vol. 31, no. 4, pp. 418–436, 2012.

[4] W. Hurst, M. Merabti, and P. Fergus, "Behavioural Observation for Critical Infrastructure Security Support," in The Seventh IEEE European Modelling Symposium (EMS2013), 2013.

[5] A. Patchimpattapong, "Development of Thailand's first nuclear power plant," in Proceedings of the International Conference on Energy and Sustainable Development: Issues and Strategies (ESD 2010), 2010, pp. 1–3.

[6] I. N. Fovino, A. Carcano, T. D. L. Murel, A. Trombetta, and M. Masera, "Modbus/DNP3 State-Based Intrusion Detection System," in 2010 24th IEEE International Conference on Advanced Information Networking and Applications, 2010, pp. 729–736.

[7] E. Knapp and J. Broad, "Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA and Other Industrial Control Systems," Syngress, Elsevier, 2011.

[8] T. Bass, "Multisensor Data Fusion for Next Generation Distributed Intrusion Detection Systems," in Proceedings of the IRIS National Symposium on Sensor and Data Fusion, 1999.

[9] E. Kuncheva, L, Combining Pattern Classifiers: Methods and Algorithms. 2004.

[10] P. Fergus, P. Cheung, A. Hussain, D. Al-Jumeily, C. Dobbins, and S. Iram, "Prediction of Preterm Deliveries from EHG Signals Using Machine Learning," PLoS One, vol. 8, no. 10, p. e77154, Oct. 2013.

[11] R. P. . Duin, P. Juszczak, P. Paclik, P. Pakalska, D. De Ridder, D. M. . Tax, and S. Verzakov, A Matlab Toolbox for Pattern Recognition, Version 4. Delft Pattern Recognition Research, 2007.

[12] F. Lotte, "Study of Electroencephalographic Signal Processing and Classification Techniques towards the use of Brain-Computer Interfaces in Virtual Reality Applications," 2009.

[13] Z. Xu;, I. King, M. R.-T. Lyu, and R. Jin, "Discriminative Semi-Supervised Feature Selection Via Manifold Regularization," IEEE Trans. Neural Networks, vol. 21, no. 7, pp. 1033–1047, 2010.

[14] N. J. Salkind, Statistics for people who (think they) hate statistics. 2008, p. 3rd ed. Sage Publications