Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 03, March -2018

Medical Claim Fraud Analysis Disclosure Trends in Cyber Security using Machine Learning Techniques

Mr.M.Suresh[1], D.kavitha[2], C.Sangheetha[3], R.Sowmya[4]

[1]*Information Technology, Manakula Vinayagar Institute of Technology*
[2]*Information Technology, Manakula Vinayagar Institute of Technology*
[3]*Information Technology, Manakula Vinayagar Institute of Technology*
[4]*Information Technology, Manakula Vinayagar Institute of Technology*

Abstract-- *Complex Big Data systems in modern organization are gradually becoming targeted by existing and emerging new threat agents. Intricate and specialized attacks will increasingly be used to enslave vulnerabilities and weaknesses. With the ever-increasing trend of cybercrime incidents happening due to these vulnerabilities, the effective vulnerability management is imperative for the modern organizations regardless of their size. However, organizations struggle to manage the sheer volume of vulnerabilities discovered on their networks. Moreover, vulnerability management tends to be more reactive in practice. Attentive statistical models, simulating anticipated volume and dependence of vulnerability disclosures, will undoubtedly provide important perception to organizations and help them become more protective in the management of cyber risks. By influencing the rich yet complex historical vulnerability data, our proposed work and conscientious framework has enabled this new capability. By utilizing this sound framework, we initiated an important study on not only handling unrelenting volatilities in the data but also further unveiling multivariate dependence structure among the different vulnerability risks. In sharp contrast to the existing studies on invariant time series, we consider the more general multivariate case striving to capture their intriguing relationships. Through our extensive empirical studies using the real world vulnerability data, we have shown that a composite model can effectively capture and preserve long-term dependency between different vulnerability and exploit disclosures. In addition, this work gives the way for further study on the random perspective of vulnerability proliferation towards building more accurate measures for better cyber risk management as a whole.*

Keywords- *Intrusion detection system, Misuse-based techniques, anomaly-based techniques, Cognitive learning, Gradient boosting tree*

I. INTRODUCTION

The term Big Data refers to the data which is large in size which is being generated across the world at an unprecedented rate. These data can be structured, unstructured, semi-structured or quazi-structured. There is a need to convert these big data into business intelligence that enterprises can be readily located. The processed and organized data leads in better decision making and an improved way to formulate for organizations regardless of their size, type, market share, customer segmentation and other categorizations.

Cyber security is the group of technologies and processes which is designed to secure the network, computers, program and data from the attacks, unauthorized access, any change, or destruction. Cyber security systems consist of network security systems and computer (host) security systems. Each of this security system has, at a minimum, a firewall, antivirus software, and an intrusion detection system (IDS). IDS help to uncover, determine, and discover unauthorized use, duplication, alteration, and destruction of information systems. The security breaches includes two type of intrusions, the external intrusions (attacks from outside the organization) and the internal intrusions (attacks from within the organization).IDS supports three types of cyber analytics: misuse-based (sometimes also called signature- based), anomaly-based, and hybrid. Misuse-based techniques are used to detect the known attacks by using the signatures of those attacks. They are effective for detecting the known type of attacks without generating an enormous number of false alarms. They require frequent manually updated database with rules and signatures of the attacks. Misuse-based techniques cannot detect novel (zero-day) attacks.

The Anomaly-based intrusion model has both the normal network and system to determine the anomalies as deviations from normal behavior. The anomaly-based intrusion has the ability to detect zero-day attacks. Another advantage is that the profiles of normal activity are customized in every system, application, or network, which is difficult for the

attackers to know which activities they can carry out to be undetected. Additionally, the data on which anomaly-based techniques alert (novel attacks) can be used to define the signatures for misuse detectors. The main disadvantage of anomaly-based techniques is the potential for high false alarm rates (FARs) because previously unseen (yet legitimate) system behaviors may be categorized as anomalies. Hybrid techniques combine misuse and anomaly detection. They are employed to raise detection rates of known intrusions and decrease the false positive (FP) rate for unknown attacks. An in-depth review of the literature did not discover many pure anomaly detection methods; most of the methods were really hybrid. Therefore, in the descriptions of ML and DM methods, the anomaly detection and hybrid methods are described together.

Another division of IDSs is based on where they look for intrusive behavior: network-based or host-based. A network- based IDS identifies intrusions by monitoring traffic through network devices. Host-based IDS monitors process and file activities related to the software environment associated with a specific host.

II. BACKGROUND

Anna L. Buczak et al,. 2, this work describes the machine learning (ML) and data mining (DM) methods for cyber analytics in support of intrusion detection. This work representing each method were identified, read, and summarized. Because data are so important in ML/DM approaches, some well-known cyber data sets used in ML/DM are described. The disadvantages of ML/DM algorithms are addressed, discussion of complexity for using ML/DM for cyber security is presented, and some recommendations on when to use a given process are provided. Bhuyan et al. [2]; instead it concentrates only on ML and DM techniques. However, in addition to the anomaly detection, signature-based and hybrid methods are depicted. The descriptions of the methods in the present survey are more in-depth than in [2]. Nguyen et al. [3] describe ML techniques for Internet traffic classification. The techniques described therein do not rely on well-known port numbers but on statistical traffic characteristics. Their survey only covers papers published in 2004 to 2007, where our survey includes more recent papers. Unlike Nguyen et al. [3], this paper presents methods that work on any type of cyber data, not only Internet Protocol (IP) flows.

Teodoro et al. [4] focus on anomaly-based network intrusion techniques. The authors present statistical, knowledge-based, and machine-learning approaches, but their study does not present a full set of state-of-the-art machine-learning methods. In contrast, this paper describes not only anomaly detection but also signature-based methods. Our paper also includes the methods for recognition of type of the attack (misuse) and for detection of an attack (intrusion). Lastly, our paper presents the full and latest list of ML/DM methods that are applied to cyber security.

Sperotto et al. [5] focus on Network Flow (Net Flow) data and point out that the packet processing may not be possible at the streaming speeds due to the amount of traffic. They describe a broad set of methods to detect anomalous traffic (possible attack) and misuse. However, unlike our paper, they do not include explanations of the technical details of the individual methods.

Wu et al. [6] focus on Computational Intelligence methods and their applications to intrusion detection. Methods such as Artificial Neural Networks (ANNs), Fuzzy Systems, Evolutionary Computation, Artificial Immune Systems, and Swarm Intelligence are described in great detail. Because only Computational Intelligence methods are described, major ML/DM methods such as clustering, decision trees, and rule mining (that this paper addresses) are not included.

There is a paradox to digital revolution trends where IT consumer's effect has brought an increased number of security threats for modern enterprises. The substantial impacts of major cyber security breaches have proliferated extensive and cross-disciplinary research work [1], [4], [5], [6], [21], [23]. A high-level risk management framework was described in [21] using cyber-risk insurance as the foundation for decision planning. An empirical study was conducted in [5] to estimate the impact of vulnerability information disclosure and availability of patches from both attackers and software vendor's perspectives. The study considered attack frequency based on the honeypot data as the crude impact measure. Their preliminary results suggested that the vulnerability disclosure increased the frequency of attacks, which also forced vendors to release patches earlier.

The paper [19] took a different stance on the substantiated economic damages because of software vulnerabilities and their disclosure policies. The authors emphasized the increasingly crucial role of understanding the emerging market structures and policy implications on the cost of vulnerabilities. A multi-disciplinary approach towards understanding and managing information security was further advocated in [4]. Our statistical framework and empirical insights essentially contribute to this line of research and strengthens cross-disciplinary understandings on the evolution of vulnerability

disclosures. Our work is also closely related to another line of research explicitly analyzing and modelling trends in vulnerabilities [10].

The patching process was investigated and linked with the life-cycle of vulnerability. The vulnerability lifecycle was formally captured by three different stages namely discovery, exploitation and patching. The paper revealed that exploits would normally become available earlier than the release of corresponding patches. The process of vulnerability discovery has been extensively studied in [20]. A more recent study in [20] applied time series techniques to build vulnerability forecasting models for five popular web browsers in cluding Chrome, Firefox, IE, Safari and Opera. The authors in [16] investigated the growth of software vulnerabilities covering a number of operating systems including Red Hat and Microsoft. The paper suggested a sigmoidal model can be used for describing the growth.

A novel measure was proposed by [23] called time between vulnerability disclosures (TBVD). It provides some important insights on the likelihood of finding zero-day vulnerability by an expert analyst within a given timeframe. The Gamma distribution was found to provide the best fit with different scale parameters depending on the software products (e.g. Linux and Windows). Most of the vulnerability models implicitly assumed the independence of vulnerability disclosures, which overlooked market-driven incentives. In addition, the proposed models only factored in mean behaviors and are limited to univariate time series, whereas our framework is more general and robust in the presence of extreme events. Traditionally, the theoretical foundation of time series modelling was deeply rooted in the statistics community, but the recent wave of new big data applications has nurtured some novel approaches. Time series shape let is such a new primitive driven by the data mining community. Given a time series with two classes, a shape let is essentially a subsequence with the most discriminative local features for separating them out. Without explicitly assuming the structure of input data (e.g. stationary), time series shapeless have found some novel applications in dealing with heterogeneous sensor data and malware detection.

III. TECHNICAL PRELIMINARIES

3.1. Cognitive learning

Cognitive functioning is defined as referring an individual's ability to process that should not enlarge on a large scale in healthy individuals. It is said to be as the capability of an individual person to perform the various mental processes mostly associated with the learning and the problem solving techniques, like verbal, words, spatial, psychomotor, and processing-speed abilities. Cognition is the things of memory, the capacity to learn new information, word, speech, and understanding of written material and data. The brain is usually has the ability to learning new skills in the previously mentioned typically in early childhood, and of developing personal thoughts and beliefs about the world. Humans when they born have a capability for cognitive function, so almost every person is capable of learning or remembering new information. However, their capability is tested by using various tests like the IQ test, psychometric test, although these have issues with accuracy value and completeness. In these tests, the individuals will be asked various sequences of questions or to perform various tasks, with each measuring a cognitive skill such as level of consciousness in work, memory management, awareness, problem-solving, motor skills, analytical abilities, or other similar concepts. Early stage of childhood is when most people are able to absorb much information and use that new information. In this period, people learn many new words, concepts, and various methods to express their thoughts.

Cognitive functioning is defined as referring to an individual's capability to process to thoughts that should not effect on a large scale in healthy individuals. It is referred as "an ability of an individual to perform the various mental process and activities most closely associated with learning and problem solving techniques". The cognitive learning system can "think like a human" with the ability to understand speech and gather, research, analyze and interpret both unstructured documents, images and videos and pre-formatted, machine-ready structured data. This work we consider the document databases and it's preprocessed and stored as an excel format.

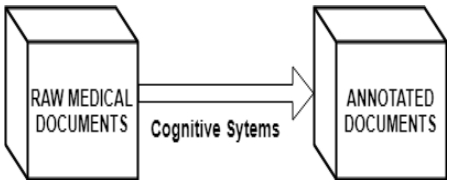

Figure 1. Cognitive Learning

3.2. Gradient boosting tree

It is an machine learning technique for regression and classification problems, which produces a prediction model in the form of an collection of weak prediction models such similar to decision trees. It builds the model in a stage-wise process like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. The gradient boosting algorithm as iterative functional of gradient descent algorithm. This algorithm optimizes a cost function over function space by iteratively choosing the function as weak hypothesis those points in the negative gradient direction. This functional gradient view of boosting has led to the development of boosting algorithms in many areas of machine learning and statistics beyond regression and classification.

Input

Training set $\{(x_i,y_i)\}n_{i=1}$, a differential loss function $L(y,F(x))$, Number of iterations M.

Gradient boosting tree algorithm

1. Initialize model with a constant value:

$$F_0(x)=\arg\min\sum_{i=1}L(y_i,\gamma)$$

2. For m=1 to M:

i) Compute so-called pseudo-residuals

$$r_{im}=-[\partial L(y_i,F(x_i))/\partial F(x_i)]_{F(x)=Fm-1(x)} \quad \text{for } i=1,2,..n$$

ii) Fit a base learner h(x) to pseudo-resuidals, ie., train it using training set $\{(x_i,r_{im})\}_{ni=1}$

iii) Compute multiplier r_m by solving the following one dimensional optimization problem

$$r_m=\arg\min\sum L(y_i,F_{m-1}(xi))+\gamma h_m(x_i))$$

iv) Update the model $F_m(x)=F_{m-1}(x)+r_m h_m(x)$

3. Output $F_m(x)$

IV. MEDICAL CLAIM FRAUD ANALYSIS

The proposed system is to detect the fraud happening in the medical insurance claim. The machine learning techniques used to detect the accuracy in the medical claim details and this accuracy value helps in finding the fraudulent data. 15% of total claims (health) are false in health care industry. India is losing approximately 600 to 800 cores annually in health care industry. The fraud detection can be done by using machine learning techniques such as, Cognitive learning system and Gradient Boosting Tree algorithm (GBT). Using cognitive learning helps to pre-process the data by ordering the data and eliminating the correlated data. Using Gradient Boosting Tree (GBT) algorithm we can optimize the loss function and make predication on weak leaner through decision tree. The detected fraudulent data is taken as the output.

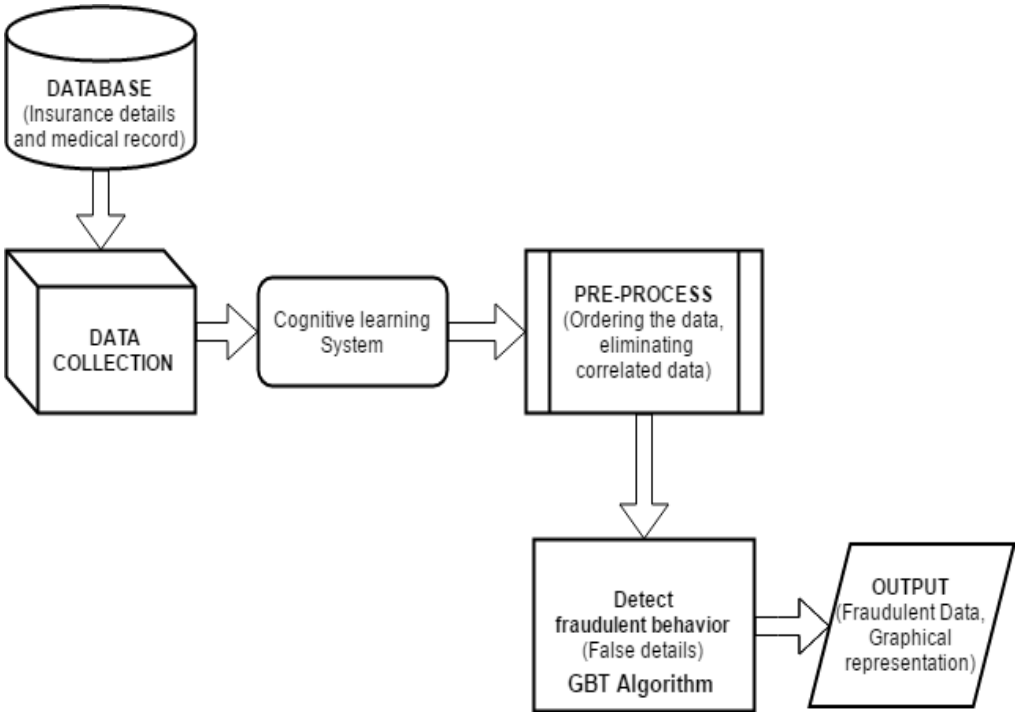

Figure 2. Medical claim fraud analysis architecture

4.1. Data Collection

The medical claim datasets are collected from the Tokyo-based Fukoku Mutual Life Insurance Co. This insurance company will slash about 30 percent of its claims payment from the department staff for IBM's Watson-powered AI system that will help manage over 1,30,000 insurance claims annually. The medical record and insurance details are collected together. These collected details are undergone under cognitive system for preprocessing.

4.2. Data Pre-processing

The pre-processing steps are: Data cleaning, Data smoothing, ordering of data, Removing Correlated data. Data cleaning is the process of detecting and correcting inaccurate records from a dataset. Data smoothing is the process of removing the noisy data from the dataset. Ordering of data is arranging of data set in a sequence for processing. Removing correlated data is the process of removing the attributes which has 95% of correlation in it. The preprocessed dataset is given under cognitive learning system. The cognitive learning system can "think like a human" with the ability to understand speech and gather, research, analyze and interpret both unstructured documents, images and videos and pre-formatted, machine-ready structured data. This work we consider the document databases and its preprocessed and stored as a excel format. Cognitive learning system uses natural language processing and machine learning techniques. It is use to sense the data between raw data and actionable data. The cognitive learning system converts raw medical documents into annotated documents.

4.3. Detect Fraudulent Behavior

A gradient boosted model is used to determine the regression and classification tree models. It is a forward-learning group method that gathers the predictive results through gradually upgraded estimations. Boosting is a workable nonlinear regression process that helps to improve the accuracy of the decision trees. By orderly applying the weak classification algorithms to the upgraded changed data, a series of decision trees are created that gives a group of weak prediction models. Since the boosting trees increases their accuracy values, it also helps to decreases time taken for processing and the human interpretability. The gradient boosting method generalizes tree boosting to minimize the issues in determining the accuracy. The gradient boosting tree algorithm produce prediction module using decision trees. The gradient boosting tree algorithm is for three purposes: Optimize loss function, prediction of weak leaner, model to add weak leaner to minimize the loss function.

4.4. Visualization and Performance Evaluation

The result is provided in graphical representation model which gives the accuracy of true and false data. Performance Evaluation is a list of performance criteria values are calculated on the basis of the label and the prediction attribute of the input Example Set. The output performance vector which has the performance criteria is defined by the performance operator (known as calculated-performance-vector). If a performance Vector was also given at the performance input port (known as input-performance-vector here), the norm of the input performance vector are also included in the output performance vector. The proposed Gradient Boosted tree is compared with Naive Bayes algorithm and normal decision tree to estimate the accuracy. If the input performance vector and the calculated performance vector both have the same criteria but with different values, then values of the calculated performance vector are delivered through the output port.

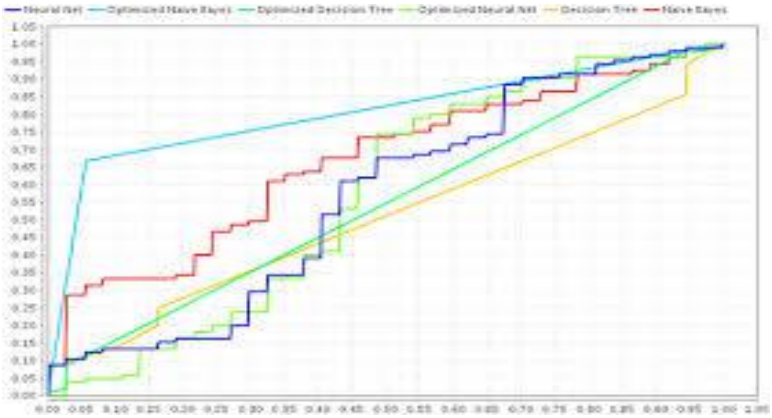

Figure 3. Resultant Graph

V. CONCLUSION

The fraud happening in the medical insurance claims creates huge loss in the health industry. Thus finding the fraud happening in medical insurance helps the health industry to save huge amount of money. The machine learning techniques such as cognitive learning and gradient boosting algorithm helps in identifying the false insurance claims. The resulted graph gives the accuracy of the insurance claim data. This accuracy values determines the false details. In future process of this work we have decided to use online data. As we taken only the offline data to detect the fraudulent insurance claim details, in future we try to take both online data and offline data which gives more accurate results when comparing to the results taken only by offline data.

VI. REFERENCES

[1] MingJianTang, MamounAlazab, *Senior Member, IEEE*, and YuxiuLuo, Big Data for Cybersecurity: Vulnerability Disclosure Trends and Dependencies,Vol.X,No.X,OCTOMBER2016

[2] M. Bhuyan, D. Bhattacharyya, and J. Kalita, "Network anomaly detec- tion: Methods, systems and tools," IEEE Commun. Surv. Tuts., vol. 16, no. 1, pp. 303–336, First Quart. 2014.

[3] T. T. T. Nguyen and G. Armitage, "A survey of techniques for inter- net traffic classification using machine learning," IEEE Commun. Surv. Tuts., vol. 10, no. 4, pp. 56–76, Fourth Quart. 2008.

[4] Axelsson S. The Base-rate fallacy and its implications for the difficulty of intrusion detection. ACM Transactions on Information and System Security 2000;3:186–205.

[5] Computer Economics, "2007 malware report: The economic impact of viruses, spyware, adware, botnets, and other malicious code," Jul. 2008. [Online]. Available: http://www.computereconomics.com

[6] M. S. Abadeh, J. Habibi, Z. Barzegar, and M. Sergi. A paral- lel genetic local search algorithm for intrusion detection in com- puter networks. Engineering Applications of Artificial Intelli- gence, 20(8):1058–1069, 2007.

[7] MamounAlazab, Profiling and classifying the behaviour of malicious codes, Journal of Systems and Software Elsevier, Vol. 100, 2015, 91-102, ISSN0164-1212.

[8] M.Alazaband R. BroadhurstSpam and criminal activity, Trends & issues in crime and criminal justice no. 526, Australian Institute of Criminology,2016

[9] M.S.Shahzad, Copula modelling of dependencein multivariate time series, International Journal of Forecasting, Elsevier, 2015, Vul. 31, No. 3,815-833.

[10] P. Johnson, D. Gorton, R. Lagerstro¨ m and M. Ekstedt, Time between vulnerability disclosures: A measure of software product vulnerability, Computers &Security, Elsevier, Vol. 62, 2016,278-295.

[11] Common Taxonomy for the National Network of CSIRTS (Includes Legal Framework),EUROPOL-Eur. Cybercrime Centre (EC3), The Hague, The Netherlands, Jul.2016.

[12] M. Bozorgi, L.K. Saul, S. Savage and G.M. Voelker, Beyond Heuris- tics: Learning to Classify Vulnerabilities and Predict Exploits, Pro- ceedingsof the 16th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, 2010,105-114.

[13] H. Yuanrong, C. Xi, I.Bose, and W. Qiu-Hong, Cybercrime Enforcement: A Comparative Study of US, UK, China, and Other European Countries.2014.

[14] R. M. Kowalski, G. W.Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyber bullying research among youth," Psychol. Bull., vol. 140, no. 4, pp. 1073–1137,2014.

[15] Communication From the Commission to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions „Towards a Comprehensive European Framework for Online Gambling", Eur. Commission ,Brussels, Belgium, 2012, accessed on Mar. 9, 2016.

[16] Gaffney J, Ulvila J. Evaluation of intrusion detectors: a decision theory approach. IEEE Symposium on Security and Privacy 2001:50–61.

[17] Heckerman D. A tutorial on learning with Bayesian networks. Microsoft Research; 1995. Technical Report MSRTR-95-06.

[18] Kabiri P, Ghorbani AA. Research in intrusion detection and response – a survey. International Journal of Network Security 2005;1(2):84–102.

[19] Kruegel C, Valeur F, Vigna G, Kemmerer R. Statetul intrusion detection for high-speed networks. IEEE Symposium on Security and Privacy 2002:285–94.

[20] Kruegel C., Mutz D., Robertson W., Valeur F. Bayesian event classification for intrusion detection. In: Proceedings of the 19th Annual Computer Security Applications Conference; 2003.

[19] M. Amini and R. Jalili. Network-based intrusion detection using unsupervised adaptive resonance theory. In Proceedings of the 4th Conference on Engineering of Intelligent Systems (EIS '04), Madeira, Portugal, 2004.

[20] M. Amini, R. Jalili, and H. R. Shahriari. RT-UNNID: A practical solution to real-time network-based intrusion detection using unsu- pervised neural networks. Computers & Security, 25(6):459–468, 2006.

[21] J. An, G. Yue, F. Yu, and R. Li. Intrusion detection based on fuzzy neural networks. In J. Wang, ZhangYi, J. M. Zurada, B.-L. Lu, and

[22] H. Yin, editors, Advances in Neural Networks - Third Interna- tional Symposium on Neural Networks (ISNN '06), volume 3973 of Lecture Notes in Computer Science, pages 231–239. Springer Berlin / Heidelberg, 2006.

[23] K. P. Anchor, P. Williams, G. Gunsch, and G. Lamont. The com- puter defense immune system: current and future research in intru- sion detection. In D. B. Fogel, M. A. El-Sharkawi, X. Yao, G. Green- wood, H. Iba, P. Marrow, and M. Shackleton, editors, Proceedings of the IEEE Congress on Evolutionary Computation (CEC '02), volume 2, pages 1027–1032, Honolulu, HI, USA, 12-17 May 2002. IEEE Press.

