Data Mining and Intrusion Detection Systems

Zibusiso Dewa and Leandros A. Maglaras
School of Computer Science and Informatics
De Montfort University, Leicester, UK

Abstract—**The rapid evolution of technology and the increased connectivity among its components, imposes new cyber-security challenges. To tackle this growing trend in computer attacks and respond threats, industry professionals and academics are joining forces in order to build Intrusion Detection Systems (IDS) that combine high accuracy with low complexity and time efficiency. The present article gives an overview of existing Intrusion Detection Systems (IDS) along with their main principles. Also this article argues whether data mining and its core feature which is knowledge discovery can help in creating Data mining based IDSs that can achieve higher accuracy to novel types of intrusion and demonstrate more robust behaviour compared to traditional IDSs.**

Keywords—(Intrusion Detection; NSL–KDD; Machine Learning; Datasets; Classifiers; Feature Selection; Waikato Environment for Knowledge Analysis; Anomaly detection; Misuse detection; Data mining)

I. Introduction

As the years have passed by computer attacks have become less glamorous. Just having a computer or local network connected to the internet, heightens the risk of having perpetrators try to break in, installation of malicious tools and programs, and possibly systems that target machines on the internet in an attempt to remotely control them. The (GOA) team categorised the attacks encountered in 2014 discovering that 25% of the attacks where non-cyber threats followed by scan/probes/attempted access 19% and policy violation 17% [1]. This data is further acknowledged by the annual FBI/CSI survey which discovered that though virus based attacks occurred more frequently, attacks based on unauthorised access and denial of service attacks both internally as well as externally, increased drastically.

Recent exploits also suggest that the more sensitive the information that is held is, the higher the probability of being a target. Several Retailers, banks, public utilities and organizations have lost millions of customer data to attackers, losing money and damaging their brand image [2]. In some cases attackers steal sensitive information and attempt to blackmail companies by threatening to sell it to third parties [5]. In the second quarter of 2014, Code Spaces (source code company) was forced out of business after attackers deleted its client databases and backups. JP Morgan, Americas' largest bank, suffered a cyber-attack in 2014 that impacted 76 million members [3]. In 2014, Benesse, A Japanese Education Company for children suffered a major breach whereby a disgruntled former employee of a third-party partner disclosed up to 28 million customer accounts to advertisers [4]. Most notably the "Sony Pictures hack" best displayed how significant a companies' losses are in the aftermath of a security breach. The network servers were temporarily shut down due to the hack [4]. Cybersecurity experts estimate that Sony lost up to $100 million [5] [6]. Other companies under the Sony blanket fell victim to attacks [7]. To tackle this growing trend in computer attacks and respond threat, industry professionals and academics are joining forces in a bid to develop systems that monitor network traffic activity raising alerts for unpermitted activities. These systems are best described as Intrusion Detection Systems.

II. Intrusion Detection Systems

A. *Definition of Intrusion Detection*

Heady et al. [8] describes an intrusion as a set of actions that make attempts to challenge the integrity, discretion or accessibility of a resource. Generally the practice of intrusion detection involves the tracking of important events which take place in a computer system and analysing them in order to detect the potential presence of intrusions [9]. Alessandri gives a more comprehensive definition of intrusion detection, describing it as a collection of practices and mechanisms used to detect errors that may lead to security failure with the use of anomaly and misuse detection and by diagnosing intrusions and attacks [8].

Correspondingly it may be added that an intrusion detection system is the practical implementation of intrusion detection principles and mechanisms over a network [8]. This is combination of software and/or hardware components that run on a host machine monitoring the activities of users and programs searching for possible insider threats on the host device and also inspecting network traffic of networks that are connected to the host, looking for outsider threats [8]. The objective of an IDS is to alert administrators of suspicious activities and in some cases even attempt to circumvent the attacks. The practices employed in IDSs' do differ from other security techniques such as firewalls, access control or encryption which aim to secure the computer system. With this being identified however it is strongly recommended that these security practices are used in conjunction with one another as this reinforces defence of a system and ensures that a much larger scope of a system is protected [9].

B. *History of Intrusion Detection*

Originally, Intrusion Detection (ID) was conducted manually by system administration. They were tasked with thoroughly monitoring each activity on a console identifying any anomalies. This early form of ID proved ineffective due to the errors it produced. Automated log file readers where then developed allowing quick searching for irregularities and unauthorised personnel [8]. It is worth noting that early versions of ID were owned by few organisations, computing was not a widespread practice and the technological computing age had not been born [8]. The introduction of audit logs helped manifesting ID into a forensic technique; whereby administration collated information and only identified issues after incidents had already occurred and not during the process of an attack. Before the 90s' Intrusion detection was a form of post analysis, analysis of intrusions and changes in system structure were only identified long after the actual event. The processes were tedious, slow time consuming and presented potential of human error due to heavy involvement [11]. During the '80 to 90s' research was carried out in a bid to strengthen existing ID software. Some suggest that the breakthrough came in the 90s' as a result of the Intrusion Detection System proposed by Denning [12]. Researchers developed an IDS that reviewed audit data as it was produced. This advancement spawned the first version of real time IDSs' allowing for attack pre-emption through methods of real time response [10]. As the world entered the technological age, the market demand for IT security increased and IDS were further developed and made available to large organisations. New features were developed such as various new alert methods, updates to attack pattern definitions, dedicated user friendly interfaces and prevention techniques that automatically stopped attacks when identified [11].

With the focuses now shifting toward enhancing security measures, newer attack techniques continued to spawn from every corner of the web; most notably the Millennium bug and Morris worm. Due to this it became apparent to developers that in an ever changing environment one must always seek to improve and stay ahead as threats become more diverse in their methods to find new ways to penetrate systems.

C. *Understanding taxonomy of IDS*

A general definition of Taxonomy is the practice or principle of classification [9]. Taxonomy may serve several purposes in design. Firstly, it can describe the current global situation, assisting in refining complex situations and presenting it in a clearer measured approach (Description). Additionally using taxonomy to classify a number of objects, enables identification of missing objects early in design, which allows users to exploit the predictive qualities of a good taxonomy (Prediction). Lastly, a good taxonomy presents users with ideas, further explaining observed current occurrences (Explanation).

There have been many taxonomies presented for IDSs' and ID technologies, dating back as far as 1999. The first real recognised IDS taxonomy seems to be the one proposed by Debar et al. [18]. Since then many more taxonomies have been published, most notably is the one proposed by Axelsson [9] followed by another proposed by Halme and Bauer [19]. The identified taxonomies can be used in order to illustrate general relationships in IDSs'. Figure 1 which is illustrated below displays the revised version of the taxonomy previously proposed by Debar et al. in 1999, this version features additional criteria for classifications [39].

Debar explains the significance of understanding the system before creating an IDS and expanding on various mechanisms used in IDS to enable a structured approach in design [39].

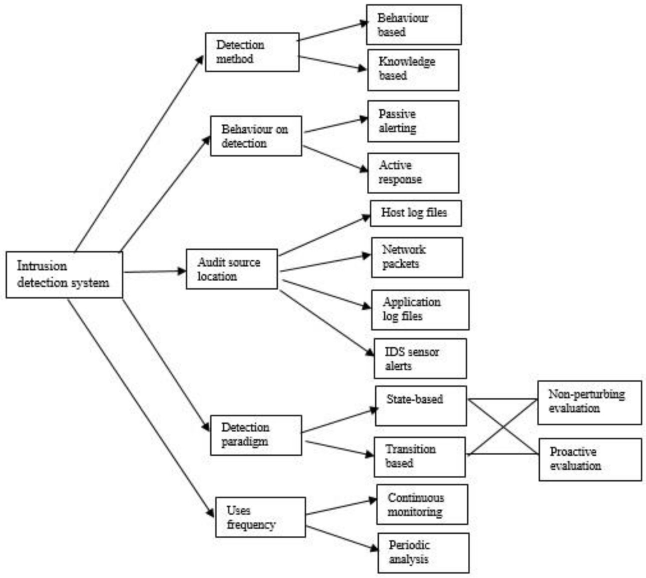

Fig. 1. Updated IDS taxonomy

It is believed that following this practice results in development of efficient Intrusion Detection Systems. There is no clear indication as to whether this is a justified approach to creating IDSs', however the sheer sum of researchers who incorporate this technique into their design may serve as evidence, suggesting the importance of taxonomy. The Halme and Bauer [19] taxonomy named "A taxonomy of Anti-Intrusion Techniques", focuses on unexplored methods in combating intrusion activities, it focusses on this rather than dealing with IDSs'.

The taxonomy reveals six anti-intrusion approaches prevention, pre-emption, deflection, deterrence, detection, and/or autonomously countered. Axelssons' taxonomy propose an enhanced approach, as it deals solely with the IDSs [14]. This commences with classification of the section principles, followed by the operational tasks of the IDS. The taxonomy aims to analyse current IDSs, which consequently allows progress in researching the chosen field enabling categorisation to help aid enhance knowledge of the field. In Figure 1 the taxonomy follows a series of steps, firstly the classification of the detection principle and then certain operational aspects of the intrusion detection system. In Figure 1, initiation starts through first identifying the different types of intrusive behaviours generated by an intruder. Further progressing to questioning suitable practices on how to observe these intrusions (intrusion sources), the repercussion of doing so and finally the outcome and decision.

D. *Intrusion Detection Sytsem Catergories*

By further analysing Axelssons taxonomy, two methods of IDS may be discovered when viewed from another perspective as illustrated in Table 1 [17]. The first is building a taxonomy using principals of IDSs', where categorisation is based upon the following detection methods; Anomaly detection, Signature detection and Hybrid/compound detection. Following on from this the author establishes IDS classifications, based upon system characteristics, time of detection and response to detecting intrusions as presented in Table 1.

TABLE I. GENERAL IDS TAXONOMY

Anomaly	Self learning	Non-time series
		Time series
	Programmed	Descriptive stats
		Default deny
Signature	Programmed	State modelling
		Expert system
		String matching
		Simple rule-based
Signature inspired	Self learning	automatic feature selection

III. IDS PRINCIPLES

An intrusion detection system functions by determining whether a set of actions can be deemed as intrusion on a basis of one or more models of intrusion. A model describes a list of states or actions as good or bad (potential intrusion) [20]. These ID methods can be implemented into two different system categorisations. Anomaly detection system which is identifies network traffic behaviour and misuse detection system which bases its detection on signatures or pattern matching, also described as knowledge based.

A. *Anomaly Detection*

An anomaly detection based system uses the normal profile of a system or user to determine its decision making process[10]. Development begins at the point at which the detector forms a judgement on behaviour that constitutes to normal for the observed object in question (application, user, resource usage etc.) and then a percentage of this activity may be flagged as suspicious and a preserved action is then taken [14]. This type of system is suited for the detection of previously unknown attacks as it detects most intrusions without having acquired prior knowledge of the intrusion [20,40,41]. However, issues still remain as it fails to expand on details surrounding the particular intrusion, (fault diagnosis), in addition this system is noted to return high false positive rates [10]. These detection methods can be used to define signatures for misuse detectors and when merged with a misuse system can form a hybrid system [12].

1) Self Learning systems

When using self learing IDS, no underlying information is made available to them. The IDS typically learns by observing the traffic and creating a model for each system's fundamental process [14]. Self-learning IDS may use "non-time series" or "times series" approaches to formulate a model for the normal behaviour of the system. Time series is the more complex technique as it requires time to be taken into account; examples of these techniques the Hidden Markov Model (HMM) and Artificial Neural Network (ANN).

2) Programmed

In this classification, an expert programmes the system to detect certain irregular events. The system does not learn on its own and requires an expert to form the normal behaviour profiles of the system; then later deciding actions to be considered abnormal that trigger an alert. Thus, this is the user of the system who defines the normal operation. Axelsson states that programmed IDSs use two techniques firstly, descriptive statistics, whereby the system builds a profile of normal statistical behaviour by the parameters of the system and then gathering descriptive statistics on selection of parameters. Secondly is default deny, this is where the class plainly states each status whereby the system functions in or around a realm deemed safe and secure; deviations from this state will flag as intrusions [14].

Halme and Bauer [19] further categorise anomaly detection systems through system specification and profiling. Determining the system components and the behaviours to capture and monitor, allows for different classes to appear. Classes such as Threshold Monitoring, user profiling, group profiling, static work profiling and adaptive rule based profiling are discussed.

B. *Misuse Detection*

Authors in [11] and [12] demonstrate Misuse Detection Systems commonly referred to as signature based detection systems, because they work by using patterns of recognised attacks or known critical points in a system to find and match known intrusions. The decisions made, are formed on the basis of knowledge of a model, dealing with the intrusion process and what is to be tracked in the observed system. Misuse detection offers greater accuracy and can efficiently detect variations of recognised attacks. Furthermore, such IDS also offer more meaningful intrusion diagnostics when an alarm is triggere by detailing diagnostic information about the cause of an alarm [8]. However these detection systems also have some pitfalls as they lacks ability in detecting new attacks and signatures that are unavailable [10]. Contrary to this it must be noted that many commercial systems employ misuse detection systems such as Cisco which employs knowledge-based systems.

1) Programmed

In this approach the system is programmed from the offset with a clear decision rule set. This rule set, offers simplicity as it contains coding of expected responses in the event of an intrusion. Another variation of programmed misuse IDS is the State Modelling method, where the intrusion is coded as a number of different states; each of them must exist within the observed space to be determined as an intrusion. These methods are in fact time series models. Two subclasses exist within the method, State transition which states the chain that is negotiated from beginning to end and petri-net. Both subclasses establish a petri-net, the structure of this system is similar to that of a tree and several states can be satisfied in any order no matter where they may occur in the model.

The expert system class is an intelligent system employed to evaluate the security state of the system, when assigned rules that describe intrusive actions. Forward chaining, production tools are often used as they are appropriate when dealing with systems, where new data are continually entered into the system. String matching method is also used for matching case sensitive characters in text, which are exchanged among systems. Simple rule-based systems are simplified versions of an expert system. These tend to execute quicker as they are not as advanced [14]. To conclude, Axelsson suggests in his latest revision of his taxonomy that current research must be redeployed in studying the effectiveness of intrusion detection and how to handle attacks against intrusion detection systems themselves.

Countless research has been conducted in an attempt to answer this question; with many quality contributions noted in this area, however there is still plenty of space to improve areas in IDS development and fulfilling the proposal made by Axelsson. Based on expert opinions there is a general consensus regarding the current state of network IDS. Many organizations are opting into purchasing signature based intrusion detection systems, due to the fact that they require less supervision, offer more automation and consume less time in setting features; therefore there is a belief that chances of human error are reduced. Furthermore, its widely stated that the majority of these organisations will employ IDSs' that are not suited to their system needs as they simply pick the biggest brands, which may offer simplicity but on the same time they are left without an understanding of how to use these systems. Many experts state that issues still remain in identifying new forms of intrusion and in order to stay ahead, the cyber security industry must continue to develop IDS and organisations train key staff on how to use these devices rather than relying solely on automation.

IV. Data Mining

This section of this article will aim to combat the concerns raised by Axelsson as well as other experts, a review of existing literature is undertaken with surveys on current datasets in IDS, effective use of classifier algorithms, identifying relevant fields for feature selection and suitable ranking systems to test performance characteristics when a test is undertaken.

Computers have been identified as one of the sole orchestrators in building a platform to move technology enhancements. This has also had an impact on network traffic monitoring solutions, with huge volumes of data generated, some of which being heterogeneous and from different origins and travelling across devices at high speeds. Subsequently all of these factors makes it difficult to produce accurate analysis of data in a timely manor [12]. Data mining is identified as a solution to handling the analysis of data due to its adaptability and validity and it is now used extensively used for network security purposes [13].

Authors in [32] describe how intrusion detection systems categorise network traffic as either an anomaly or normal. Data mining is employed into an intrusion detection system as a method of extracting the huge volumes of data that exist in network traffic for further analysis [14]. As an application of machine learning, data mining holds a very significant position in intrusion detection, presenting methods of predicting future patterns based on past experiences [15].

However, in the same way significant researchers suggest that major challenges lie within intrusion detection and evidence demonstrates the difficulties in current data mining tools, such as high False Alarm Ratio (FAR), and low Detection Ratio (DR). Further development have been suggested in current data mining tools as questions have been raised about the quality of tools implemented [16]. Witten states that to be able to answer current questions surrounding data mining one must grasp the concept of learning while working with data mining [19]. Witten defines Data mining as a topic that involves learning in a practical and non-theoretical sense. In a way the author signifies that learning within data mining is essentially steered by the ability to think whilst also having purpose. The researcher concludes that learning without purpose is simply training and not a practice of data mining.

Knowledge discovery in databases (KDD) is a term most frequently used interchangeably with data mining and it is defined as the application of a scientific method to data mining [18]. The typical process over KDD model includes methodologies for extracting, preparing data and making decisions about actions once mining has taken place. Maimon identifies that the process commonly has 4 to 12 steps [18].

Step 1, The idea is to begin with the goal identification task, understanding a particular domain, anguish knowledge discovery is required. Step 2, The next stage is to identify a target dataset, an initial set of data for analysis. Step 3, the pre-processing of data, the use of available resources to move noisy data and decide how to deal with redundant data values and so on. Step 4, the transformation of data, the addition or elimination of attributes and instances from the target datasets, decisions on matters of normalisation, combination and smoothing of data. Step 5, here the most appropriate features for representing data is built using data algorithms. Step 6, analysis of the outputs from the previous step, determining the usefulness of the discovery and deciding whether to change steps prior to 5, possibly using different attributes to achieve different results. Step 7, if the knowledge is deemed suited it is applied and incorporated directly as a solution to the problem [18].

The following scenario described by Witten gives an understanding of the possible data mining application into current practices; the combination of knowledge discovery stages identified by Maimon enable illustration of the learning definition that Witten explains about data mining [18]. The first three stages of knowledge discovery can be noted within the scenario. The problem domain is firstly identified, subsequently data for analysis and characteristics is specified then desired features are defined. Witten illustrates this concept by using the following example; human in vitro fertilisation involves collecting many eggs from a womans ovaries, once fertilised several embryos are produced. Some are selected and transferred to the womans uterus. Challenges here lay in the subject of identifying the best embryos' to use, and the most likely to survive. Selection is based upon the 60 features of an embryo, characterisations such as follicle, sperm sample, morphology etc. This large number of features creates an issue amongst embryologists in assessing them all concurrently while also correlating historical data to determine if an embryo was likely to result in a child or not. Data mining and machine learning algorithms are used to solve the identified issue. The practice of data mining has been applied to many fields, such as sales, healthcare, medical, finance, multimedia and most importantly intrusion detection [17].

It can be concluded that Data mining, offers more than simply finding data and applying algorithms over it. Both seemingly accredited the lack of efficiency identified within data mining to lack of understanding among researchers with Witten suggesting that several publications have errouneously used data mining procedures [16]. Running many algorithms over a particular dataset and writing results, claiming one method or machine learning algorithm over another with little understanding of the nature of the dataset; poor understanding of the learning algorithm and no deliberation of the statistical importance in results [16,17] are some representative paradigms.

V. Applying Data Mining Algorithms to Intrusion Detection

The growth of data mining methods has consequently brought forth a wide range of algorithms drawn from areas as pattern recognition, machine learning and database analysis. There are many types of algorithms that may be used to mine audit data. Lazar identifies data algorithms as a set of heuristics and designs between data mining models [19]. In effect, the author suggests that for a model to be formulated, the algorithms must start by analysing the data type provided, in order to find particular trends and patterns. These results of analysis are later used by the algorithm for defining optimal parameters to create the selected mining model. The parameters are applied across the dataset, together with selected patterns and detailed statistics [13].

Scholars, Manish and Hadi conducted an investigation into network traffic analysis and prediction techniques following this they established a list of currently used data mining techniques. The authors use Table 2 to present the most commonly used data mining techniques.

TABLE II. Data Mining Techniques

Data mining techniques
Clustering
Classification
Hybrid
Association
Other methods

Numerous studies indicate that classification techniques and clustering are by far the most widely used data mining techniques. The hybrid technique is considered shortly after together with the Association technique [15,13,20].

Manish and Hadi [13] stating that clustering is the process of splitting data into clusters based upon the features of the data. This clustering partitions data into groups of similar objects. Each member within the cluster is similar to one another [13]. Witten adds by describing clustering as an unsupervised learning method primarily used when training the normality model for anomaly detection and situations where little knowledge of the attack class is required while training. Wahono further expands on the functionality of techniques by expanding on methods of grouping together into clusters using distance functions. In addition several clustering, classification algorithms are identified, however the most widely used seemed to be the k-means classification [20].

Classification and prediction, as seen in Table 2 are described as the most popular mining techniques; allowing for extraction of models, describing important data classes and aiding in predicting future trends [13]. Wahano expands on this stating that it classifies data into metrics-based classification such as normal or abnormal in intrusion detection systems. More importantly he states that classification maps data items to one of many predefined categories. The classifier's output can be used to predict a model that may forecast future trends, when sufficient normal and abnormal behaviour is gathered in audit data. The classification algorithm may be able to predict new unseen data classifying it by using pre-existing information. In addition, the author identifies some of the most widely used approaches in data classification; Bayesian classification, decision tree induction, neural network and statistical learning [20].

Manish and Hadi, briefly summarise the contrast between clustering algorithms and classification algorithms; classification algorithms require knowledge in both normal and known attack data in order to separate classes during detection [13]. Manish and Hadi determines that the Association technique discovers anomalies by using association rule algorithms, suggesting that the best applications for using the technique is; finding as many related defects to the detected defect within data, evaluating results during inspections and analysing reasons for anomalies recurring within data. Wahano argues that this technique is better suited to handling forensic analysis and not real time attacks, suggesting that the process is time consuming and would not benefit network analysis when scanning the system. The hybrid technique is a combination of two or more approaches for analysis of network traffic. The hybrid model achieves good results in the analysis of network traffic. We present various hybrid model techniques that are investigated by researchers for network traffic analysis.

Wahano briefly describes hybrid models stating that they are the combination of two or more approaches that may be utilised when analysing network traffic. Further highlighting that these approaches are generally new developments in data mining, offering new solutions in network analysis, however they are under-utilised and many of these techniques are difficult to be implemented. Notably, Haratian states that rapid development in data mining has introduced a wide range of attribute-value conditions [22]. These values often occur together in given sets of data and may help develop an understanding when determining relationships between the fields in database records. The author uses a market transaction basket example to illustrate how the method groups relevant data. Through analysis of the confidence and rule support figures within data, a system can judge the possibility of an action occurring, for example the age of a customer and the income they produce per year, the research suggest how likely it is for the individual to purchase a dvd player with the use of statistical analysis to understand the associations that may occur.

VI. Model of IDS

This section will examine the components required to construct a model of IDS. First it uncovers datasets, and the finally classifiers. It is of utmost importance to understand the contents of a dataset, and the purpose of the attacks featured in order to help design and build appropriate tools [16].

VII. Intrusion Detection Data Sets

Amudha et al. [21] presents a diagram which illustrates the knowledge discovery process (Figure 2). This process notably aids in creation of a model for intrusion detection system. At the first stage the dataset is chosen and pre-processing techniques are used to clean data.

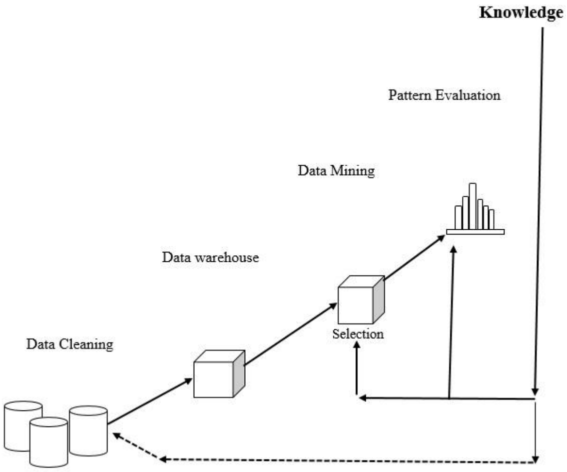

Fig. 2. Steps in knowledge discovery

Extensive research shows that the following datasets were commonly used for investigations by recognising that the KDD dataset, is widely known among scholars in the network analysis community, offering a large scope of records in the previous DARPA dataset, with up to 4,900,000 training instances, 41 features, 24 training and testing attacks with a further 14 types. The dataset offers more information compared to the DARPA dataset. However, Bajaj and Arora state that the KDD dataset is outdated, suggesting that the NSL-KDD dataset is the most suited for current network analysis. They state that KDD 99 dataset suffers with redundant data which often lead to biased detection of attacks, highlighting more frequency in DOS and probe attack. This lead to failures in classifying features appropriately and most records cannot be classified and are misrepresented in most of the cases. The author states that if the KDD, datasets are used, investigations will likely present results that do not represent real network situations.

Neethu also states that the KDD dataset is outdated and due to technological advances can no longer provide accuracy for evaluation. However the lack of alternatives is the reason the dataset is still in use. Bajaj and Arora identify The NSL-KDD as the most appropriate network analysis dataset as it manages to eliminate identical records in testing data and redundant instances in training data. This means that, classifiers are more likely to be able to categorise data if records are matched correctly. The author states that this factor can affect the accuracy of a classifier producing better results than the KDD Dataset. In contrast, Neethu believes that the NSL-KDD dataset contains some redundant instances and the dataset cannot be used for the correct training of models.

Neethu suggests that the most approipriate dataset is the ADFA-LD dataset, introduced in 2013. Normal training data holds up to 833 traces and also 10 attacks in attack data, this dataset is believed to have a closer resemblance between attack and normal data, than KDD security datasets. However the scholar does state that many researchers have struggled to utilise the dataset as features are not best described and it is difficult for many developers to understand its funcionality. This adds confusion when creating effective models for intrusion detection. The author, however believes that further experimentation in the dataset is required to demonstrate its effectiveness in network analysis [23] [24].

Amongst the majority of researchers, there seemed to be no preference in use of datasets within intrusion detection. There is a vast amount of datasets available on the internet and as Witten states the choice of a specific dataset should be accompanied by an understanding of the problem that one wishes to address, an understanding of classifiers that may be implemented over it and a way to effectively carryout an investigation [16].

VIII. Extracting Features

Feature reduction and selection are commonly used in current intrusion detection publication. The methods are often used interchangeably to indicate specific points. Feature reduction/extraction is the process of finding new subspaces with less dimensions than the original feature space [34]. Further expansion to explain feature selection is that the features established by feature selection must always be a subset of an original set of features, in contrast feature reduction reduces dimensions combining the linear combination of the original set and establishing new synthetic features, the least important features are discarded [34]. Independent component analysis

A. *Feature Reduction*

Feature reduction is finding a new subspace which has less dimensions than the original feature space. The following commonly used methods for feature reduction are presented, Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Uncorrelated Linear Discriminant Analysis (ULDA), Independent component analysis (ICA) [5]. With regards to the pre-processing technique of Feature selection, it seems that many invocative hybrid techniques are being tested in response to the current outcry to improve issues of efficiency and accuracy discovered in current Datasets. Wahba et al. [25] developed a multiclass classification to aid in producing modern intrusion detection models with improved efficiency and accuracy. The aim was to merge different classifiers together to produce better results. The authors highlight the significant benefits of using multiclass IDS backed with recent research suggestions and investigations. The author further identifies that the overall performance of the model degrades during implementation. This is caused by attempts to fully merge classification patterns where features do not match, this results with redundancy in data [25]. The author proposes a technique to reduce irrelevant features and improve the performance of the model.

Tesfahun and Bhaskari [26] reveal that due to the erratic characteristics of intrusion detection, there is huge disproportion between the classes in the NSL-KDD dataset making it difficult, to apply machine learning effectively in the area of intrusion detection. The synthetic minority over sampling technique is applied to the training data set in an attempt to deal with class imbalance. Feature selection method based on information gain is used to create the reduced feature subset. The classifier used as random forests algorithm with SMOTE and IG. The author states random forests classifiers were chosen over decision tree classifiers due to the fact that random forests algorithms can run on large datasets, have the ability to handle nominal data and do not over fit, classification is conducted through votes of string. The NSL-KDD datasets is chosen as benchmark due to research acknowledging that, KDD 99 calls a huge number of redundant records. Therefore preventing the identification of minority classes such as U2R and R2L. NSL-KDD dataset was used for test purposes following investigations the number of features were reduced to 22, the detection rate for U2R increased and time to build reduced.

Dhafian et al. [27] Investigates current literature surrounding classification techniques and methods of intrusion detection. Reviewing current IDS approaches using the following datasets; DARPA, KDD 99, NSL, KDD, Kyoto 2006 + and CAIDA. Findings suggest that NSL–KDD performed best overall once trained against specified classifiers. The authors conclude by suggesting consideration must be taken during developing of classification techniques identifying optimal dataset that is a rich the recent attacks and which features are selected without confusion, unnecessary overhead and time-consuming selection.

Desale and Ade [28] propose a Genetic Algorithm based Feature Selection Approach for Effective Intrusion Detection System. The genetic algorithm is used to search method when selecting features from the full NSL-KDD dataset. The mathematical intersection principle is used selecting features that appear in every experiment. The investigation is carried out on both test and training data set, proposed approach is measured against the popular approaches, namely the Correlation Feature Selection (CFS), Information gain (IG), Correlation Attribute Eval (CAE) and the effect on the performance of the Naive Bayes and J48T algorithm classifiers are measured. The resulting information determined that the proposed model selected the minimum features from the dataset, which improved the classifier, accuracy of the Naive Bayes classifier while also reducing time, complexity.

Chabathula et al. [29] propose Principal Component Analysis approach Using Machine for feature reduction and effective selection. The authors adopts a different approach to network analysis in order to reduce data features. Header fields of incoming packets are analysed in vectors, these serve as import for the PCA algorithm. The System is designed in two phases. The training and testing phase are conducted over the full NSL-KDD dataset. PCA produces features which are deemed weak. The test records are compared with the base profile during training phase and then the confusion matrix is used over the classification algorithms to determine performance. Results in the accuracy of detection time of each algorithm is measured. The following classifiers are tested upon within the investigation SVM, KNN, J48 tree, random Forest tree, classification, Adaboost, nearest neighbour, naive Bayes classifier and voting features classification. Two experiments take place using each classifier, one with PCA applied to it and without PCA. If an attack is present an alarm is sounded. Results display performance of classifiers in terms of detection time, testing accuracy and speed. They indicate that tree-based classifiers such as J48 and Random Forest notably outperformed other classifiers in accuracy and speed. With linear growth up until the 18th dimension when it steadily increased. Voting features classification had a lower detection rate with PCA than without, the naive Bayes problemistic classifier displayed better results with PCA than without it. Detection rates noted in SVM, KNN, J48 tree, random forest tree, Adaboost, nearest neighbour, were almost similar in both with PCA and without PCA.

Chauhan and Bahl [30] propose performance Analysis of Dimension Reduction Techniques with Classifier Combination for Intrusion Detection System. A Review of current dimension reduction techniques, search methods, attribute evaluators and classifiers was conducted. Different combinations feature selection and feature classification algorithms were applied to the datasets to detect intrusion. Results show that there was an increase in classification accuracy from 52% to 96% of PCA analysis. Scholars suggest that classification with a good accuracy results in a reduction in completion time and effective outputs.

B. *Feature Selection*

Feature selection is described as a method whereby specific features are selected from a set of features, which have a high discrimination capability between class labels. It is required to reduce the irrelevant features by eliminating them by using some methods. The method may be wrapper based or filter based or combination of them [29]. Feature selection is used to select sets with minimum length while ensuring maximum classification accuracy attributes of most importance and increase performance speed by reducing the irrelevant features through elimination [29].

Ganapathy et al. [31] review current feature selection and classification techniques for IDS, surveying intelligent techniques for developing IDSs then developing a new IDS using two proposed algorithms. Once an understanding was developed, a test scenario was formulated whereby a subset of KDD consisting of 10% of its records was used to test the algorithms. The researchers found that Modified Mutual Feature Based are more flexible during feature selection than Gradual feature removal as they use mutual information rather than relying on predetermined clustering when removing features. CRF based feature selection methods can handle uncertainty effectively and the wrapper based methods uses a decision tree to remove subsets of features. Analysis of these methods highlights that mutual information and information gain ratio provide the best methods of feature selection as they can be used to perform tuple reduction and attribute selection. After reviewing linear programming methods for detecting U2R attacks. They determine that a layer approach adopted through a Least Squares Support vector Machine will offers solutions of a linear equation to conquer trivial SVM methods, simplifying, detection of normal or attack data and improving accuracy and detection time. Furthermore they add that a neuro-tree classifier may also be used as it offers effective classification when optimal features are provided to it. A classification technique named IREMSVM is then proposed from current intelligent agent -based multiclass SVM, the new feature uses information gain ratio and attribute selection to effectively compute the feature set in a timely manner. Two new algorithms, the IREMSVM, IAEMSVM and SVM were then tested against a full KDD dataset using all feature and one with selected features. Accuracy was analysed through comparing Probe, DOS and other attacks as the proposed algorithm uses constrain checking in classification. The conclusion was that Accuracy was higher in IREMSVM than in SVM or IAEMSVM.

Zargari and Voorhis [32] examine significant features in anomaly detection systems with an aim to apply them to data mining techniques. Identifying some current challenges of obtaining a comprehensive feature set and establishing a system that eradicates redundant and recurring data from the KDD 99 dataset while also keeping the feature set to a minimal size. Rough set theory dependency was used to identify the most discriminating features of each class. Feature 21 and 22 in the KDD dataset were found not to have any significance in intrusion detection (FTP session and hot login). A further five features were identified to have a small significance in intrusion detection. These were su attempted, number of file creation operations, is guest login, dst host rerror rate. To produce a distinctive report finding the features and characteristics of the intrusion detection that offers more attacks and distribution of attacks as compared to KDD dataset. Corrected KDD-dataset was used, in order to discover the features and characteristics of the intrusions plus, whether anomaly detection can be improved by using this dataset from a statistical point of view. It is important to mention that different to other studies, the Corrected KDD-dataset was analysed here instead of the KDD-dataset. The Corrected KDD-dataset contains more attacks and the distribution of attacks is different comparing to the distribution of attacks in the KDD-dataset. A subset of features was later proposed to help decrease dimensions of KDD and compare to subset features through data mining techniques. The proposed features were later tested on NSL –KDD and demonstrated higher detection rates for proposed features. The work may require live analysis before we can be sure that it would function correctly.

Aparicio-Navarro et al. [33] finds three scenarios in which correctly labelled datasets are required. Stating that when using unsupervised IDS there is a need for labelled datasets to be trained. When the nature of an analysed data set must be recognised to evaluate efficiency of IDS when detecting an intrusion. Finally using feature selection that only works if processed datasets are labelled. Finding the flaws in current practices of labelling datasets states that collection of labelled datasets from real-world networks is impossible as many datasets are labelled through off-line forensic analysis, which is impractical as it does not allow real-time implementation. The author, develops an approach that automatically generates labelled traffic datasets with unsupervised anomaly based IDS. The resulting labelled dataset are subsets of the original unlabelled dataset. The remaining dataset may contain valuable information and so is kept so an administrator may add or remove data as they see fit. The newly labels dataset are processed using genetic algorithm (GA) approach, that performs the feature selection. This GA is implemented to automatically provide the metrics to generate appropriate intrusion detection results while reducing the risk of misclassification. The Ffitness is identified to have an important role in implementation of the technique allowing fine tuning of outcomes through maximising DR minimising FPR number of metrics or other parameters. Relating to Yasmen and Jyoti the author acknowledges the measurement of efficiency of IDS, also stating important aspects of evaluation lie in the DR, FPR, FNR, which provide quantifiable evidence of effectiveness in ID. For an IDS to be evaluated in those terms, the nature of analysed information must be normal. On the other hand, knowing the nature of this analysed information is not required during the intrusion detection process and is only necessary in evaluating IDS efficiency.

Zhang and Wang [34] investigate current feature selection techniques in a bid to build an effective solution. The author uses for commonly used in selection methods to elucidate features that are least important in the dataset. IG, ReliefF, GR, ChiSquare are used over the full NSL-KDD dataset to identify 20 important attributes. These 20 features are applied to the proposed Bayesian network classification model for feature selection. The proposed model calculates the most useful features from the 20 further reducing the figure. A investigation is conducted to compare and accuracy of commonly used feature selection methods as well as the results of the proposed method. The authors conduct tests over the benchmark dataset NSL-KDD with all of its records in the training set and 10 fold cross validation for training and testing. A comparison between the filter and wrapper approach are made with the wrapper being chosen to the fact that it evaluate features by performance of the classifier. The subset results in the best performing classifiers being selected. Bayesian networks are accepted of a model to the widespread belief that it is suited in working under uncertainty.

Relan and Patil [35] study effective ways of feature selection, comparing two algorithms. The C4.5 , decision tree algorithm and the C4.5 algorithms pruning and testing proposed features over the KDD 99 and NSL kDD dataset to test and train the classifier algorithm. After identifying decision tree technique as a logical method with advantages and extracting features and rules. The authors identify that to train machine algorithms historical data is necessary, and since the KDD dataset holds up to 4,94,020 records they choose NSL KDD to train and test the dataset. The authors decide on reducing the features in the training data. The author acknowledges the use of C4.5, which uses information gain and splitting criteria to deal with continuous and discrete attributes and uses pruning to ensure that over fitting the decision tree does not occur. When training the authors alternated between selecting the 10 fold cross validation technique and the partitioning methods, randomly choosing the percentage of the dataset to use for training and testing. The two classification algorithms were tested upon the two different datasets c4.5 decision tree and c4.5 with pruning only discrete value attributes were considered during classification. The time required for testing was less than the training classifier and the results showed that the c4.5 with pruning algorithm performed better than the C4.5 algorithm. After training the classifier KDDCup 99 and NSL-KDD test data is tested against both c4.5 decision tree and c4.5 decision tree with pruning. During C4.5 decision tree algorithm the author considered all the 41 attributes of KDDCup 99 dataset while at the time of C4.5 with pruning only discrete value attributes like protocol_type, Service, flag, land, logged_in, is_host_login, is_guest_login and class are considered during classification. The performance is measured in terms of classifier accuracy, percentage of true positive, percentage of false positive and time required for testing. The time required for testing is less as compared to training the classifier. The results generated by both the classifiers are compared with each other as shown below.

IX. Data Mining Tools

There are various datamining tools that may be used to conduct investigations. In Zupans publication the author highlights the significance in choosing the correct data mining, by first theoretically matching the correct classifier technique to the field where it would work [36]. Understanding characteristics such as fundamentals operations, the way it works phases of work then, doing the same with current data mining tools to identify patterns and ensure appropriate functionality and features are installed to create a model. Once this has been completed the author can begin to use the chosen methods in the medical biometrics field. The investigation helped illustrate the effectiveness of The WEKA tool in diagnosing Leukaemia. Many researchers share the same views expressed by the author, with several data mining investigations carried out notably through the WEKA tool rather than other popular data mining tools, such as rapid miner, KNIME and so on[36].

Most notably, they are a number of publications that study algorithms on the DARPA dataset, the KDD cup dataset and the NSL – KDD dataset which used the WEKA environment. When testing upon Knowledge discovery databases it is noted by Jagtap that WEKA supports many standard data mining tasks such as data integration, selection transformation and evaluation [37]. Knowledge extraction is a key process for businesses tools such as WEKA that allow for clear, operations or all levels of users is significant [37]. Furthermore, the report by the Pharmine company states that WEKA achieved the highest performance in accuracy amongst the data mining tools [36] [37]. Weka also offers some functionality that other tools do not, such as the ability to run up to six classifiers on all datasets, handling multi-class datasets which other tools continue to struggle with tools [36] [37]. Frank states that this effectively means that complex critical algorithms may be experimented on in this complimenting environment allowing innovative flexible research while decreasing technological limitations in research [38].

X. Conclusion

In recent years there has been a large interest in identifying the best feature set attributes for IDS classifiers. With the growing number of intrusions reported there is cause for creating accurate IDSs with low percentages of false positives. Data mining based IDSs have demonstrated higher accuracy, to novel types of intrusion and robust behaviour. Furthermore, it has been noted that intrusion detection must keep up with the sheer size, speed and dynamics which modern networks are expected to operate on. Many have tackled the issue in research papers with various datasets, namely the Knowledge Database Discovery cup 99. Fewer attempts have been made to adapt this test into the NSL-KDD dataset. NSL-KDD is noted to be one of the best representations of network traffic in the current time frame. The need for more sophisticated and adaptive IDS systems remain and Industry Professionals and academics continue to develop and present novel methods that can cope with new more sophisticated attacks.

References

[1] United States of Americe. US Govermnent Accountability Office (2015) (2015) Report on Cyber Security - Actions needed to Address Challenges facing Federal Systems. Washington: GAO-15-573T.

[2] L. Morgan (2014), List of Cyber Attacks and Data Breaches in 2014. IT Governance, 23 Dec. Available from: http://www.itgovernance.co.uk/blog/list-of-the-hacks-and-breaches-in-2014/ [Accessed on 13 November 2015]

[3] M. Watson, (2014). JP Morgan suffers data breach affecting 76 million customers. IT Governance, 23 Dec. Available from: http://www.itgovernanceusa.com/blog/jp-morgan-suffers-data-breach-affecting-76-million-customers/ [Accessed on 14 November 2015]

[4] “Report and response regarding Leakage of Customers’ personal Information.” (10 September 2014). Last accessed on 17 February 2015, http://blog.benesse.ne.jp/bh/en/ir_news/m/2014/09/10/uploads/news_20140910_en.pdf.

[5] S Tobak. (18 December 2014). Fox Business. “3 Revelations from the Sony Hack.” Last accessed on 29 January 2015, http://www.foxbusiness.com/technology/2014/12/18/revelations-from-sony-hack/.

[6] A, Peterson. (5 December 2014). The Washington Post. “Why it’s so hard to calculate the cost of the Sony Pictures hack.” Last accessed on 29 January 2015, http://www.washingtonpost.com/blogs/the-switch/wp/2014/12/05/why-its-so-hard-to-calculate-the-cost-of-the-sony-pictures-hack/.

[7] Trend Micro Incorporated, (22 December 2014). Simply Security. “The Reality of the Sony Pictures Breach.” Last accessed on 29 January 2015, http://blog.trendmicro.com/reality-sony-pictures-breach/.

[8] R. Heady, G.F. Luger, A. Maccabe and M. Servilla, “The architecture of a Network Level Intrusion Detection System,” Department of Computer Science, College of Engineering, University of New Mexico, 1990, pp. 1-17.

[9] R. Bace and P. Mell, “NIST Special Publication on Intrusion Detection Systems,” Booz-Allen and Hamilton inc, Mclean VA, 2001, pp. 5-22.

[10] R.A. Kemmerer and G. Vigna, “Intrusion Detection : A brief History and Overview,” Computer, 2002 [supplement to security and privacy magazine], pp. 27-30.

[11] J. Allen, A. Christie, W. Fithen, J. McHugh and J. Pickel, “State of the practice of intrusion detection technologies,” vol. CMU/SEI-99-TR-028, Canergie-Mellon Univ Pittsburgh PA Software Engineering Institute, 2000 , pp. 3-23.

[12] Abhaya, K. Kumar, R. Jha and S. Afroz, “Data Mining Techniques for Intrusion Detection: A Review,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, June 2014, pp. 6938- 6941.

[13] R.J. Manish, and H.T. Hadi, “A review of network traffic analysis and prediction techniques” unpublished.

[14] S. Choudhury and A.Bhowal,“Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection.” In: Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, IEEE, 2015, pp. 89-95.

[15] S.B. Kotsiantis, I.D. Zaharakis and P.E. Pintelas, “Machine Learning: a Review of Classification and combining Techniques,” Artificial Intelligence Review, vol. 26, November 2006, pp. 159-190.

[16] I. Witten, E. Frank and M. Hall, “Data mining: Practical Machine Learning Tools and Techniques.” 3rd ed., Burlington, MA: Morgan Kaufmann, 2011, pp 5-40.

[17] M. Masud, L. Khan and B. Thuraisingham, "*Data mining tools for malware detection*," Boca Raton, FL: CRC Press, Febuary 2012, pp. 15-38.

[18] O. Maimon and L. Rokach, (2010) "Data Mining and Knowledge Discovery Handbook," 2nd ed., New York: Springer Science & Business Media, 2010, pp. 1-43.

[19] A. Lazar, "Heuristic Knowledge Discovery for Archaeological Data using Genetic Algorithms and Rough Sets1," In Heuristic and Optimization for Knowledge Discovery, H. Abbass, C. Newton, & R. Sarker, Eds. Hershey, PA: Idea Group Publishing*, 2002 ,* pp. 263-278.

[20] R.S. Wahono, "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks," *In Journal of Software Engineering,* vol. 1, April 2015, pp. 1-16.

[21] P. Amudha, S.Karthik and S.Sivakumari, "Classification Techniques for Intrusion Detection–An Overview," *In International Journal of Computer Applications,* vol. 76, 2013, pp. 33-40.

[22] M.H. Haratian, "An Architectural Design for a Hybrid Intrusion Detection System for Database," unpublished.

[23] A.I. Abubakar, H. Chiroma, A.S. Muaz and L.B. Ila, "A Review of the Advances in Cyber Security Benchmark Datasets for evaluating Data-Driven Based Intrusion Detection Systems," *In Procedia Computer Science,* vol. 62, 2015, pp. 221-227.

[24] K. Bajaj and A. Arora, "Dimension Reduction in Intrusion Detection Features using Discriminative Machine Learning Approach." *In IJCSI International Journal of Computer Science Issues,* vol. 10, 2013, pp. 324-328.

[25] Y. Wahba, E. Elsalamouny and G. Eltaweel, "Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction." In *IJCSI International Journal of Computer Science Issues,* vol. 12, issue 3, May 2015, pp. 355-368.

[26] A. Tesfahun and D.L. Bhaskari, "Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction," In: *Cloud & Ubiquitous Computing & Emerging Technologies, 2013 International Conference*, 2013, pp. 127-132.

[27] B. Dhafian, I. Ahmad and A. AL-Ghamid, "An Overview of the current Classification Techniques in Intrusion Detection," in *International Conference Security and Management,* 2015, pp. 82-88 .

[28] K.S. Desale and R. Ade "Genetic Algorithm based Feature Selection Approach for Effective Intrusion Detection System," in *Computer Communication and Informatics, 3rd International Conference,* 2015, pp. 1-6.

[29] K.J. Chabathula, C.D. Jaidhar, M.A. Ajay Kumara, "Comparative Study of Principal Component Analysis based Intrusion Detection approach using Machine Learning Algorithms," in *Signal processing Communication and Networking, 3rd International Conference*, 2015, pp. 1-6.

[30] H. Chauhan, V. Kumar, S. Pundir and E.S. Pilli, "A Comparative Study of Classification Techniques for Intrusion Detection," in *Computational and Business Intelligence, International Symposium*, 2013, pp. 40-43. IEEE.

[31] Ganapathy. S et al., "Intelligent Feature Selection and Classification Techniques for Intrusion Detection in Networks: A survey," *in EURASIP Journal on Wireless Communications and Networking,* vol. 1, issue 271, 2013, pp. 1-16.

[32] S. Zargari and D. Voorhris, "Feature Selection in the Corrected KDD-dataset," In *Emerging Intelligent Data and Web Technologies, 3rd International Conference*, 2012, pp. 174-180.

[33] F. Aparicio-Navarro, K.G. Kyriakopoulos and D.J. Parish, "Automatic Dataset labelling and Feature Selection for Intrusion Detection Systems," in *IEEE Military Communications Conference, 2014, pp. 46 - 51.*

[34] F. Zhang and D. Wang, "An Effective Feature Selection Approach for Network Intrusion Detection," in *Networking, Architecture and Storage, IEEE Eighth International Conference*, 2013, pp. 307-311.

[35] N.G. Relan and D.R. Patil, "Implementation of Network Intrusion Detection System using Variant of Decision Tree Algorithm," in *Nascent Technologies in the Engineering Field, International Conference,IEEE, 2015*, pp. 1-5.

[36] S.K. David, A.T. Saeb and K. Al Rubeaan, "Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics," *in Computer Engineering and Intelligent Systems,* vol. 4, no. 13, 2013, pp. 28-38.

[37] S.B. Jagtap, "Census Data Mining and Data Analysis using WEKA," *In* International Conference in Emerging Trends in Science, Technology and Management, 2013, pp. 35 – 40.

[38] E, Frank, M. Hall, L. Trigg, G. Holmes and I.H. Witten, "Data mining in Bioinformatics using Weka*. Bioinformatics", Oxford:England,* vol. 20, 2004, pp. 2479-2481.

[39] H. Debar, M. Dacier and A.Wespi, "A revised Taxonomy for Intrusion Detection Systems," in Annales des T´el´ecommunications, vol. 55, issue 7, 2000, pp. 361–378.

[40] Leandros A. Maglaras, Jianmin Jiang, Tiago Cruz, "Integrated OCSVM mechanism for intrusion detection in SCADA systems", IET Electronics Letters, Volume 50, issue 25, December 2014, p 1935-1936

[41] Maglaras, Leandros, and Jianmin Jiang. "Intrusion detection in SCADA systems using machine learning techniques." Science and Information Conference (SAI), 2014. IEEE, 2014.