Data Mining Based Algorithms for Intrusion Detection Systems

IJOTM
ISSN 2518-8623

Volume 3. Issue II
p. 10, Dec 2018
http://jotm.utamu.ac.ug
email: ijotm@utamu.ac.ug

Anthony Arthur Ashaba
aashaba@gmail.com
Uganda Technology and Management University

Drake Patrick Mirembe
College of Computing and Information Sciences, Makerere University

Abstract

With the tremendous increase in usage of network based services. Security has remained one challenging area for networking experts. There are various security technologies that help fight the inevitable network and security attacks; they have been so vulnerable to exploitations from internal threats. This led to the development of Intrusion Detection Systems (IDS) to complement on the existing methods. Responding to and evaluating IDS alerts is labor intensive requiring vast human resource. Data Mining provides invaluable method to analyze large volume of historical computer systems data, identify patterns, trends and evaluate the behavior of threats and potential vulnerabilities and classify traffic as normal or anomalous. There are many data mining algorithms in use such as rule based approaches, Bayesian networks, Support Vector Machines (SVM) and so on. However, the performance of these algorithms is affected when no optimized features are provided. This leads to high systems processing costs and reduced performance.
This paper shows a comparative study on the various data mining techniques.

Key words: *Data mining, Intrusion detection system, Clustering, classification*

Introduction

The internet has become a part of daily life and an essential tool today. The need for an increase in security of network systems is getting more and more important day by day because of the many risks or threats associated with using the internet. There are various strategies and mechanisms that have been applied which provide security to some extent but are too static to give an effective protection [16, 36]. Even though these mechanisms provide security, they have failed to detect intrusions [6]. Deploying an Intrusion Detection System (IDS) helps increase visibility and control within a corporate computing environment.

The effectiveness of IDS depends on the capability to detect any abnormal activity in the target system. As such IDS examine all data features to detect intrusion or misuse patterns. However Intrusion Detection systems to generate a large volume of alerts which is unmanageable and overwhelming to the human analyst. Most of these alerts are false positives [1, 14, 21, 26, 27, 35, 44, 45, 48]. Data mining techniques make it possible to search large amounts of data for characteristic rules and patterns.

With the increased use of Intrusion Detection Systems as an integrated part of a security system, the challenge is that it generates a huge amount of alarms and most of them are false positive alarms. It's not possible to build a completely secure system. Northcutt and Novak [30] explain that most of the current IDS have very high rate of false positives as they cannot make wise decisions on whether the traffic is harmful or innocuous. The act of not detecting an intrusion when the observed event is illegal. If an attack occurs during the training period for establishing the base line data, this intrusive behavior will be established as part of the normal baseline. This affects signature detection. The percentages of intrusions that can be detected are kind of low meaning the patterns of known intrusions do not work well after they are developed for a particular environment and once configurations have been made for a particular network they may not work well for another because they will all have different traffic patterns.

The IDSs are not effective as we hope they are because we need to study the network system, operating systems and the attack methods that are launched against our networks. Data mining is becoming an integral part of current IDS because it empowers it to search large quantity of data for distinctive rules and patterns. The idea of applying data mining to Intrusion detection systems is to maximize the effectiveness in identifying attacks thereby helping to construct more secure information systems. Different data mining techniques like clustering, classification, association rule,and outlier detection techniques are helping the various aspects of intrusion data analysis [23]. The advantage of applying Data Mining technology to the Intrusion Detection System lies in its ability of mining the succinct and precise characters of intrusions in the system from large quantities of information automatically. It can solve the problem of difficulties in picking-up rules and in coding of the traditional Intrusion Detection system.

Many intrusion detection systems have been constructed by manual and ad hoc means. These systems have been designed and implemented based on the system builders' knowledge of a computer system and their understanding of known intrusions. As a result, the effectiveness and adaptability of the intrusion detection systems are limited in the face of new computing environments or newly invented attack methods [22]. Experts first analyze and categorize attack scenarios and system vulnerabilities, and hand-code the corresponding rules and patterns for misuse detection. An IDS often stores a database of known attack signatures and compare patterns of activity, traffic or behavior it sees in the data its monitoring against those signatures to recognize when to close. Originally IDSs consisted of a manual search for anomalies [4], however many IDS tools now store a detected event in a log which is reviewed at a later date by the administrator. These alerts need to be evaluated by security analysts before any further investigation in order to take appropriate action against attacks but manually reviewing these logs is difficult, error prone and time consuming and ignoring them may lead to successful attacks [2]. Human labeling of audit information is tedious, expensive and time consuming [9].

The following are measures to evaluate the efficiency of an Intrusion Detection System [7].

- Accuracy:-deals with the proper detection of attacks and the absence of false alarms. Inaccuracy occurs when an IDS flags a legitimate action as anomalous or intrusive.
- Completeness:-indicates sensitivity of IDS. Incompleteness occurs when the IDS fails to detect an attack.
- Performance:-indicates the rate at which audit events are processed. If the performance of the IDS is poor, then real time detection is not possible.
- Timeliness:-implies that the IDS's response or the reaction to an attack should be sooner.
- Fault-tolerance:-implies that the IDS should itself be resistant to attacks.

Host-based & Network Based IDSs

Host-based IDSs are installed on computer hosts of the network to help monitor the events occurring within that host only for suspicious activity. They get audit data from host audit trails and monitor activities such as integrity of system, file changes, host based network traffic, and system. If there is any unlawful change or movement, it informs the central management server [42]. Host based detection systems monitor data, files and operating system processes that will potentially be targets of attack. They can access system information, generating more accurate alerts and more detailed logs and they are also useful because they can detect encrypted attacks by checking traffic before being sent or just received. They cause substantial overhead in the process of securing individual hosts on the network. Examples include Snort, Dragon Squire, Emerald eXpert-BSM, NFR, Intruder Alert, etc.

Network-based IDSs are placed at strategic points within the network to monitor incoming and outgoing traffic of all devices on the network. This involves placing a set of traffic sensors within the network which perform local analysis and detection and report suspicious events to a central location. They can detect distributed alerts and have a low cost of implementation however they are weak against DoS attacks and have high requirements on computing performance to scan every packets. If an intruder can bypass it, then all systems within the network would be affected. Examples include Network Flight Recorder (NFR), Cisco Secure (formerly NetRanger), Hogwash, Dragon, ETrust.

Anomaly detection Vs Misuse detection

Anomaly detection is the process of comparing definitions of what activity is considered normal against observed events to identify significant deviations [4]. First we create a baseline profile of the normal system, or program activity. Any activity that deviates from the baseline is treated as a possible intrusion. You assume you do not know the attack method all you know is the normal behavior. The idea here is build a set of normal profiles that characterize what is normal. You then observe the run time system activities and if they deviate from the normal then you can conclude that there is something wrong. Anomaly detection has the capability to detect insider attacks for instance when a user misuses their accounts and accesses information outside the user's profile then anomaly detection generates an alarm [32], they are also useful when it comes to detecting new threats or different versions of known execution without prior knowledge of intrusion [49],[5], [25], [24], [32]. However a critical issue for anomaly detection is the high percentage of false alarms which make it difficult to associate specific alarms with the events that triggered them [32], [19]. Another drawback of anomaly is that the system must go through a training period in which appropriate user profiles are created by defining "normal" traffic profiles. The creation of an inappropriate normal traffic profile can lead to poor performance [32].

Misuse detection also known as signature based detection, is the most popular commercial type of IDS and attempts to model abnormal behavior or signatures of known attacks [38]. It uses knowledge of known attacks, exploits and vulnerabilities to look for matching patterns in network traffic or system events [39]. This operates by comparing observed events against predefined signatures in order to identify possible unwanted traffic [4], [29]. The accuracy of such systems is considered to be very good because they tend to have a low rate of false positive alarms. It is very effective at detecting known attacks but largely ineffective at detecting previously unknown threats. However there are difficulties in updating information on new attacks thereby making them unable to detect newly invented attacks. It requires a manual update of new types of attacks discovered and the human expert has to perform such task, which is time consuming. Signatures can easily be escaped with morphs of known attacks related to their knowledge database. Snort is the most popular signature based light weight network IDS and it analyses application layer of network traffic to detect specific patterns of well known attacks such as buffer overflow, portscan etc [37] Bro [34], Haystack [40].

Data Mining with IDS

Across a wide variety of fields, data is collected and accumulated at a dramatic pace. There is urgent need for a new generation of computational techniques and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of data. Bramer [3] stated that data mining is a tool that predicts future behaviors and trends. Data mining tools predict future trends and behaviors which help organizations to make proactive knowledge driven decisions.

Kalarani and Brunda [18] define data mining as the process of discovering interesting patterns or knowledge from large amounts of data. The type of interestingness could be frequency, rarity, correlation, length of occurrence, consistency, repeating / periodicity, "abnormal" behavior and so on. This interestingness is information which is not very obvious, something that is not visible directly. Fayyad, Piatetsky-Shapiro, and Smyth [11] indicate that these interesting patterns can be used to make predictions. The goal of data mining is to extract information from the dataset and change it into an understandable structure.

data + interestingness criteria = hidden pattern

The Data mining models are of two types as:

- Predictive model builds models that can learn from past data, predict future or unknown values of variables or patterns based on known data. Examples include KNN, Naïve Bayes, SVM, Networks, decision trees.
- Descriptive model analyses given datasets to identify novel or interesting or useful patterns/rules/ trends that can describe the dataset. Examples include K-means, sequence mining.

Applying data mining techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., on network traffic data is a promising solution that helps improve IDS. Data mining techniques play a vital role in intrusion detection by analyzing the large volumes of network data and classify it as normal or anomalous [10] and this is because of its ability to extract attributes from the data and the rule [31].

How data mining might contribute to intrusion detection:

- Remove normal activity from alarm data to allow analysts to focus on real attacks
- Identify false alarm generators
- Find anomalous activity that uncovers a real attack
- Identify long, ongoing patterns (different IP address, same activity)

To accomplish these tasks, data miners use one or more of the following techniques:

- Data summarization with statistics, including finding outliers
- Visualization present a graphical summary of the data
- Clustering of the data into natural categories
- Association rule discovery defining normal activity and enabling the discovery of anomalies
- Classification predicts the category to which a particular record belongs.

Classification

Classification is also known as supervised learning, is used to determine the predetermined output. It predicts the target class for each item. It assigns the data into target classes [15]. Classification algorithms require knowledge in both normal and known attack data in order to separate classes during detection [17].

Classification involves finding rules that partition data into disjoint groups. Most popular algorithms for classification in data mining are Rule based methods, decision trees, Bayes classifier, K-Nearest Neighbor, Neural networks, Support Vector Machine (SVM).

The classification process is as follows:

1. It accepts collection of items as input
2. Maps the items into predefined groups or classes defined by some attributes
3. After mapping, it outputs a classifier that can accurately predict the class to which a new item belongs

Clustering

Since the network data is too huge, labeling of each and every instances or data points in classification is expensive and time consuming [8]. Joshi and Hadi [17] stated that clustering is the process of splitting data into clusters based upon the features of data. Clustering is an unsupervised machine learning mechanism for discovering patterns in unlabeled data. It is used to label data and assign it into clusters where each cluster consists of members that are quite similar and members from different clusters are different from each other. clustering can be applied on both anomaly detection and misuse detection.

The common approach of all clustering techniques is to find cluster centers that will represent each cluster. A cluster center is a way to tell where the heart of each cluster is located, so that later when presented with an input vector, the system can tell which cluster this vector belongs to by measuring a similarity metric between the input vector and all the cluster centers, and determining which cluster is the nearest or most similar one [13], [47]. An important advantage of using clustering is the ability to find new attacks not seen before meaning that attack types with unknown pattern signatures can be detected. Clustering results can also assist the network security expert with labeling network traffic records as normal or intrusive. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction.

Another reason for clustering is to discover relevance knowledge in data. The disadvantage here is if the data is uniformly distributed, trying to find clusters of data will fail, or will lead to artificially introduced partitions. Another problem that may arise is the overlapping of data groups. Overlapping groupings sometimes reduce the efficiency of the clustering method, and this reduction is proportional to the amount of overlap between groupings [13].

Association Rule searches a frequently occurring item set from a large data set.

From the three data mining techniques discussed above clustering is widely used for intrusion detection because it does not require the use of a labeled dataset for training, so no manual classification of training needs to be done and there is no need to be aware of new types of intrusions for the system to be able to detect them [10].

Table 1: Comparison of Data Mining Techniques for Intrusion Detection Systems. [10], [12], [20], [28], [33], [39], [41], [51], [52]

Technique	Advantages	Disadvantages
Support Vector Machine	• highly accurate • less prone to over fitting than other methods • able to model complex non linear decision boundaries	• complex computation • large memory usage • the speed in both training and testing is slow

Decision Tree	• Can handle high dimensional data • Fast in classifying unknown records • Good for small size trees • Able to process both numerical and categorical data • Representation is easy to understand	• Output attribute must be categorical • Limited to one output attribute • Decision tree algorithms are unstable • Trees created from numeric datasets can be complex
Neural Network (ANN)	• High tolerance to noisy data • Availability of multiple training algorithms • It is able to implicitly detect complex nonlinear relationships between dependent and independent variables. • Requires less formal statistical training • Good for continuous data • Able to classify unknown pattern • Even if an element of the neural network fails, it can continue without any problem due to their parallel nature. • Learns and does not need to be reprogrammed.	• Complex computation • Requires long training time • Prone to over fitting • High processing time is required for • Large neural networks
Naive Bayes	• Reveal high accuracy and speed when applied to large databases. • Low computation complexity • It is easy to implement. It requires a small amount of training data to estimate parameters • Handle both continuous and discrete data • Simple computation • Not sensitive to irrelevant features	• Lack of available probability data. • It is assumed that the data attributes are conditionally independent • Large memory usage • Slow in classification testing data • Not good for high dimensional data
K-Nearest Neighbor	• It is analytically tractable • It lends itself very easily to parallel implementations • Uses highly adaptive behavior information • Simple and easy to implement	• It has large storage requirements • It is highly susceptible to the curse of dimensionality • Slow in classifying test tuples
K-means	• Easy to implement • Fast and simple • Good for large data	• Sensitive to initialization • Cannot measure the quality of clusters • Not robust to noise or outliers • Need to find out proper number
Genetic Algorithm	• has better efficiency • used to select best features for detection • Genetic algorithm based systems can be re-trained easily. It improves its possibility to add new rules and evolve intrusion detection system. • It can solve the problems with multiple solutions • It can easily transferred to	• Complex method • Mutation rate is high. • It does not have constant optimization response time
Fuzzy C-Means Clustering	• Has better robustness • Items can fit in more than one cluster	• Its performance depends on the initial number of clusters • Long time complexity
Agglomerative Clustering	• Efficiency, scalability and energy saving • Low communication overhead	• High computation complexity • Low detection accuracy • Dependence of survival score determined

Association rules	• used to detect known attack signature or relevant attacks in misuse detection	• it cannot be used for totally unknown attacks • it requires more number of database scans to generate rules
Hybrid techniques	• It is easy to implement. • It is able to compute more sets of frequent items.	• computational cost is high
Apriori	• It is easy to implement. • It is able to compute more sets of frequent items.	• It can be very slow due to the generation of large number of candidate itemsets. • It needs several scans of the dataset. • It consumes a lot of memory and hence it is suitable generally only for datasets small in size. • It could produce duplicates in the process of Candidate generation.
C4.5	• It generates classifier models which can be easily interpreted. • It can handle both continuous • and categorical values. • It can handle missing data.	• It is vulnerable to outliers. • It overfits training instances with rare features especially noisy data.
CART	• It can handle both numerical and categorical values. • It can handle outlier data. • It can identify the variables which are the most significant.	• It may produce unstable decision Trees insignificant change of training Instances may result in change in the trees. • It splits on one variable only.
Fuzzy logic	• Effective, especially against port scans and probes. • is simple and flexible • Its ability to model complex systems makes it a valid alternative in the computer security field to analyze continuous sources of data and even unknown or imprecise processes.	• High resource consumption Involved. • Reduced, relevant rule subset identification and dynamic rule updation at runtime is a difficult task. • Difficult to develop a model from a fuzzy system • Before operational it requires more fine tuning and simulation

DATASETS used with IDS

IDS algorithms need training dataset to properly function, while the research on the data used for training and testing the detection model is equally of prime concern, better data quality can improve offline intrusion detection.
The KDD Cup 1999 dataset is one of the most commonly used dataset for intrusion detection evaluation [43]. It is the most comprehensive dataset that is still valid and applied to compare and measure the performance of IDSs.

However KDDCUP'99 dataset [46] suffers from limitations due to duplication, which leads to the biasing in detection of attacks which are more frequent in data set like DOS and PROBE attacks [50]. Some researchers had used NSL-KDD [46] dataset which is the duplicates removed and size reduced version of KDD Cup '99 dataset but all the experiments are done only on anomaly detection model.

Conclusion

Anomaly detection is very intolerant to errors in that the system becomes unusable because it generates high false alarms. This could be attributed to lack of training data. It is a challenge filtering out all data that is abnormal when building the detection model. At the same time determining what is normal and abnormal is complex. What may be normal in one environment could be anomalous in another. The biggest task is evaluating

the accuracy of the system and yet devising a sound evaluation scheme is more difficult than building the system itself. Because of poor evaluation, the IDS is not able to look for specific information thereby generating high false alarms and a longer computation time.

In light of the above, data mining techniques detect the hidden and related data making IDSs effective in reducing false alarms, adaptive, dynamic and requiring minimal human intervention. From Table 1, it is difficult to choose a particular method over others while implementing an IDS because each of the approaches has its own advantages and disadvantages. But it can be concluded that if more than one approach is integrated with the IDS helps detect attacks with high accuracy.

References

[1] O. Abouabdalla, H. El-Taj, A. Manasrah, and S. Ramadass, "False positive reduction in intrusion detection system: A survey," in Broadband Network & Multimedia Technology, 2009. IC-BNMT'09. 2nd IEEE International Conference on. IEEE, 2009, pp. 463–466.
[2] K. Alsubhi, I. Aib, and R. Boutaba, "Fuzmet: A fuzzy-logic based alert prioritization engine for intrusion detection systems," International Journal of Network Management, vol. 22, no. 4, pp. 263–284, 2012.
[3] M. Bramer, Principles of data mining. Springer, 2007, vol. 180.
[4] N. Chakraborty, "Intrusion detection system and intrusion prevention system: A comparative study," International Journal of Computing and Business Research (IJCBR), vol. 4, no. 2, pp. 1–8, 2013.
[5] 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 (ICSTM), 2015 International Conference on. IEEE, 2015, pp. 89–95.
[6] R. K. Cunningham, R. P. Lippmann, D. J. Fried, S. L. Garfinkel, I. Graf, K. R. Kendall, S. E. Webster, D. Wyschogrod, and M. A. Zissman, "Evaluating intrusion detection systems without attacking your friends: The 1998 darpa intrusion detection evaluation," DTIC Document, Tech. Rep., 1999.
[7] H. Debar, M. Dacier, and A. Wespi, "Towards a taxonomy of intrusion-detection systems," Computer Networks, vol. 31, no. 8, pp. 805–822, 1999.
[8] D. K. Denatious and A. John, "Survey on data mining techniques to enhance intrusion detection," in Computer Communication and Informatics (ICCCI), 2012International Conference on. IEEE, 2012, pp. 1–5.
[9] M. Dhakar and A. Tiwari, "A new model for intrusion detection based on reduced error pruning technique," International Journal of Computer Network and Information Security, vol. 5, no. 11, p. 51, 2013.
[10] M. DŠsilva and D. Vora, "Comparative study of data mining techniques to enhance intrusion detection," International Journal of Engineering Research and Applications (IJERA) ISSN, pp. 2248–9622, 2013.
[11] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI magazine, vol. 17, no. 3, p. 37, 1996.
[12] S. Garg and A. K. Sharma, "Comparative analysis of various data mining techniques on educational datasets," International Journal of Computer Applications, vol. 74, no. 5, 2013.
[13] K. Hammouda and F. Karray, "A comparative study of data clustering techniques," Fakhreddine Karray University of Waterloo, Ontario, Canada, 2000.
[14] C.-Y. Ho, Y.-D. Lin, Y.-C. Lai, I.-W. Chen, F.-Y.Wang, and W.-H. Tai, "False positives and negatives from real traffic with intrusion detection/prevention systems," International Journal of Future Computer and Communication, vol. 1, no. 2, p. 87, 2012.

[15] V. Jaiganesh, P. Sumathi, and A. Vinitha, "Classification algorithms in intrusion detection system: A survey," International Journal of Computer Technology and Applications, vol. 4, no. 5, p. 746, 2013.
[16] L. Joseph and R. Sudha, "Data mining based intrusion detection."
[17] M. Joshi and T. H. Hadi, "A review of network traffic analysis and prediction techniques," arXiv preprint arXiv:1507.05722, 2015.
[18] P. Kalarani and S. S. Brunda, "A survey on efficient data mining techniques for network intrusion detection system (ids)." International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 9, pp. 8028–8031, 2014.
[19] R. A. Kemmerer and G. Vigna, "Intrusion detection: a brief history and overview," Computer, vol. 35, no. 4, pp. supl27–supl30, 2002.
[20] V. K. Kshirsagar, S. M. Tidke, and S. Vishnu, "Intrusion detection system using genetic algorithm and data mining: An overview," International Journal of Computer Science and Informatics ISSN (PRINT), vol. 2231, p. 5292, 2012.
[21] M. Kumar, M. Hanumanthappa, and T. S. Kumar, "Intrusion detection system false positive alert reduction technique," ACEEE Int. J. on Network Security, vol. 2, no. 03, 2011.
[22] W. Lee, S. J. Stolfo, and K. W. Mok, "Adaptive intrusion detection: A data mining approach," Artificial Intelligence Review, vol. 14, no. 6, pp. 533–567, 2000.
[23] C.-T. Lu, A. P. Boedihardjo, and P. Manalwar, "Exploiting efficient data mining techniques to enhance intrusion detection systems," in IRI-2005 IEEE International Conference on Information Reuse and Integration, Conf, 2005. IEEE, 2005, pp. 512–517.
[24] L. A. Maglaras and J. Jiang, "Intrusion detection in scada systems using machine learning techniques," in Science and Information Conference (SAI), 2014. IEEE, 2014, pp. 626–631.
[25] L. A. Maglaras, J. Jiang, and T. Cruz, "Integrated ocsvm mechanism for intrusion detection in scada systems," Electronics Letters, vol. 50, no. 25, pp. 1935–1936, 2014.
[26] M. V. Mahoney and P. K. Chan, "An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection," in International Workshop on Recent Advances in Intrusion Detection. Springer, 2003, pp. 220–237.
[27] A. Mokarian, A. Faraahi, and A. G. Delavar, "False positives reduction techniques in intrusion detection systems-a review," International Journal of Computer Science and Network Security (IJCSNS), vol. 13, no. 10, p. 128, 2013.
[28] M. Monshizadeh and Z. Yan, "Security related data mining," in Computer and Information Technology (CIT), 2014 IEEE International Conference On. IEEE, 2014, pp. 775–782.
[29] P. Ning and S. Jajodia, "Intrusion detection techniques," The Internet Encyclopedia, 2003.
[30] S. Northcutt and J. Novak, Network intrusion detection. Sams Publishing, 2002.
[31] M. Panda and M. R. Patra, "Ensembling rule based classifiers for detecting network intrusions," in Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on. IEEE, 2009, pp. 19–22.
[32] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Computer networks, vol. 51, no. 12, pp. 3448–3470, 2007.
[33] R. Patel, A. Thakkar, and A. Ganatra, "A survey and comparative analysis of data mining techniques for network intrusion detection systems," International Journal of Soft Computing and Engineering (IJSCE) ISSN, pp. 2231–2307, 2012.
[34] V. Paxson, "Bro: a system for detecting network intruders in real-time," Computer networks, vol. 31, no. 23, pp. 2435–2463, 1999.
[35] T. Pietraszek and A. Tanner, "Data mining and machine learningUtowards reducing false positives in intrusion detection," Information security technical report, vol. 10, no. 3, pp. 169–183, 2005.
[36] U. Ravale, N. Marathe, and P. Padiya, "Feature selection based hybrid anomaly intrusion detection system using k means and rbf kernel function," Procedia Computer Science, vol. 45, pp. 428–435, 2015.

[37] M. Roesch et al., "Snort: Lightweight intrusion detection for networks." in Lisa, vol. 99, no. 1, 1999, pp. 229–238.
[38] I. Singh, I. Shah, and P. Singh, "Comparative study of various distributed intrusion detection systems for wlan," Global Journal Of Research in Engineering, vol. 12, 2014.
[39] S. Singh et al., "A survey on intrusion detection system in data mining," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 1, no. 2, pp. 2190–2194, 2013.
[40] S. E. Smaha, "Haystack: An intrusion detection system," in Aerospace Computer Security Applications Conference, 1988., Fourth. IEEE, 1988, pp. 37–44.
[41] T. Smitha and V. Sundaram, "Comparative study of data mining algorithms for high dimensional data analysis," International Journal of Advances in Engineering & Technology, vol. 4, no. 2, p. 173, 2012.
[42] A. Subaira and P. Anitha, "Efficient classification mechanism for network intrusion detection system based on data mining techniques: a survey," in Intelligent Systems and Control (ISCO), 2014 IEEE 8th International Conference on. IEEE, 2014, pp. 274–280.
[43] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on. IEEE, 2009, pp. 1–6.
[44] K. Timm, "Strategies to reduce false positives and false negatives in nids," SecurityFocus Article, 2001.
[45] G. Tjhai, "Comprehensive approaches of intrusion detection in handling false alarm issue," in SEIN 2007: Proceedings of the Third Collaborative Research Symposium on Security, E-Learning, Internet and Networking. Lulu. com, 2007, p. 53.
[46] unb. ((accessed December 5, 2016)) Nsl-kdd data set for network-based intrusion detection systems. [Online]. Available: http://www.unb.ca/cic/research/datasets/nsl.html
[47] M. Verma, M. Srivastava, N. Chack, A. K. Diswar, and N. Gupta, "A comparative study of various clustering algorithms in data mining," International Journal of Engineering Research and Applications (IJERA), vol. 2, no. 3, pp. 1379–1384, 2012.
[48] G. J. Victor, M. S. Rao, and V. C. Venkaiah, "Intrusion detection systems-analysis and containment of false positives alerts," Int. J. Comput. Appl, vol. 5, no. 8, pp. 27–33, 2010.
[49] R. S. Wahono, "A systematic literature review of software defect prediction: Research trends, datasets, methods and frameworks," Journal of Software Engineering, vol. 1, no. 1, pp. 1–16, 2015.
[50] Y. Wang, K. Yang, X. Jing, and H. L. Jin, "Problems of kdd cup 99 dataset existed and data preprocessing," in Applied Mechanics and Materials, vol. 667. Trans Tech Publ, 2014, pp. 218–225.
[51] R. Wankhede and V. Chole, "Intrusion detection system using hybrid classification technique," International Journal of Computer Sciences and Engineering, vol. 4, no. 11, pp. 30–33, 2016.
[52] L. H. Yeo, X. Che, and S. Lakkaraju, "Modern intrusion detection systems," arXiv preprint arXiv:1708.07174, 2017.

Bio

Anthony Arthur Ashaba, received his BSc. (Second Class Upper Hons.) in Computer Science from Makerere University and an Msc. Data Communications and Software Engineering from Makerere University, Uganda. Currently, he is a part time Assistant Lecturer in the department of Computer Science, Gulu University. His research interest includes;- Networks and Computer Security, Data Mining, Machine Learning, Cloud computing and Internet of Things Security, Recommender systems and Preference analytics, and Social computing.

Drake Patrick Mirembe, holds a PhD in information systems security from Groningen University, Masters in Computer System Security from Radboud University Netherlands, and BSc Computer Science and Math from Makerere University. He works both in academia and industry. In Academia he works with Makerere University and Uganda Technology and Management University (UTAMU) as a Lecturer and Dean respectively. His research interests include; ICT4D, cyber security, innovation management, mobile and wireless technologies