Mesopotamian journal of Cybersecurity
Vol.2021, **pp.** 16–21
DOI: https://doi.org/10.58496/MJCS/2021/004; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/CyberSecurity

Research Article

# Challenges and Future Directions for Intrusion Detection Systems Based on AutoML

Zainab Ali Abbood[1],*, , Ismael Khaleel[2] , , Karan Aggarwal[3],

[1] *Electrical and Computer Engineering, Altinbas University, Istanbul, 34000, Turkey*

[2] *Modern University of Business and Science (MUBS), Lebanon*

[2] *Electronics and Communication Engineering Department, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, India*

**ARTICLE INFO**

Article History

Received 20 Dec 2020
Accepted 8 Jan 2021
Published 10 Jan 2021

Keywords

IDS

Network Security

Cyber Security

AutoML

**ABSTRACT**

Recent use of computer systems and the Internet has contributed to severe protection, privacy and confidentiality problems due to the processes involved in the electronic data transformation. Much has been done to improve the security and privacy of information systems, but these issues remain in computer systems; there is, in fact, no system in the world That is early stable. Furthermore, various network attacks develop when the signature database incorporates a new signature with irregular behaviour. With many types of attacks emerging, many techniques are being built and used in many forms of network attacks. Intrusion detection systems ( IDS) are one of those methods. This method allows the management of several network networks, cloud storage and an information system. The IDS can track and detect attacks to breach a system's security features (confidentiality, availability, and integrity). This research aims at classifying IDS based on their intended goal and to compare different types of IDS in each class.

## 1. INTRODUCTION

Recent advances in computing and networking have brought serious protection, privacy, and confidentiality issues associated with the processes involved in electronic data transformation[1]. Even while much work has gone into making computers safer and more private, there still needs to be a system that is entirely secure. In addition, there are numerous varieties of network attacks [2], which manifest themselves when a new signature exhibiting anomalous behaviour has been added to the signature database. With the emergence of numerous unique types of attacks, many approaches have been developed and deployed in a wide variety of network attack vectors. One such technique is an intrusion detection system or IDS. This approach can be used to control a plethora of networks, including cloud-based file storage and information system. The IDS can monitor and identify intrusion attempts (confidentiality, availability, and integrity). This study aims to categorize IDS according to their function and then compare several kinds of IDS within the same category. The increasing trend of massive data sharing over the network has rendered traditional intrusion detection methods ineffective. This lag in feedback strongly suggests that the attacker has sufficient time to complete their objective. Due to these factors, studies have aimed to improve detection accuracy and time (time required to detect the intrusion). These allow us to classify IDS research into two broad buckets: studies that prioritise speed of detection and those that prioritise accuracy.

*Corresponding author. Email: Zainab.abood.93@gmail.com

## 2. IDS Based on Neural Network

The neural network is among the earliest tools for detecting security breaches (NN). In [3], the authors present a new generation of intelligent IDS operating with minimal features. The suggested method extracts feature using the information gain (IG) and correlation concepts. A well-designed approach is used to integrate the features after scraping them using IG and correlation following extraction. By removing unnecessary information from the dataset in advance, the pre-processing method improves resource usage and decreases the complexity of the process in terms of both time and effort. The small datasets were used to train an artificial neural network (ANN) classification system, which was then applied to five subsets of the KDD99 dataset for evaluation. In addition, host-based IDS employing log files has been considered in terms of the GRNN and MPNN models [26]. An ANN-based classification system was developed in [4], which was tested on five distinct subsets of the KDD99 dataset after initial training on the compact datasets. In [4], the authors offer a unique instance selection algorithm based on fuzzy logic, which can be used to optimise the training data. It is clear from the outcomes that the new, improved training data set can successfully learn the boundaries between normal and abnormal class labels. Membership vectors for each training set were obtained by employing a random weight neural network (RWNN), a base classifier, in this study[5]. Using a backpropagation learning technique, a powerful Anomaly Network Intrusion Detection System (ANIDS) based on the Back Propagation Neural Network (BPNN) was developed [6]. This novel algorithm improves upon previous methods in terms of accuracy and detection rate while simultaneously reducing the occurrence of false positive alarms.

## 3. IDS BASED ON SVM

SVMs can be used for classification and regression analysis because they are supervised, learning models. Multiple studies have used the SVM-based IDS. Using support vector machines, a practical identification technique for boosting features was proposed[7]. The feature-augmented method is employed here to supply SVM classifiers with informative and high-quality training data. This not only shortens the training time for the SVM but also improves its ability to detect intrusions. GPSVM[8], a novel categorizationcategorisationlso proposed. This system utilizes the NSL-KDD dataset in an effort to speed up Anomaly detection and improve classification accuracy without automatically implementing any reduction strategy, such as feature selection or resampling. A Triadic Representation of the Tao, Sun and Suna For use in a human-controlled intelligent IDS, [9] presented a GA and SVM-based alarm intrusion detection system named FWP-SVM-GA. This study initially took advantage of the GA's population search approach and interpersonal information sharing by adjusting the GA's crossover and mutation probability. Recently, advancements have been made to increase the SVM's convergence and learning algorithm speed. Similarly, another research paper[10] developed a GA that uses Hypergraphs for intrusion detection (HG-GA). Like the SVM, this framework used a technique for picking features and adjusting parameters. Effective IDS with a high detection rate and low false positive alarm rate was created by applying a weighted objective function. The hyper-clique process in Hypergraph simplifies the GA's search for the best answer. To further optimize parameters, the particle swarm method (PSO - OCSVM) was developed[11].

## 4. DECISION TREE-BASED IDS

One machine learning classification algorithm is the decision tree (DT). The DT has been used in a wide variety of IDS-related investigations. A novel approach was provided for compressing the representation spaces of individual KDD 99 ID datasets before implementing specific machine learning algorithms. [12]. The experimental results indicate that this method may guarantee a high detection rate and a low rate of false positives across a variety of datasets while simultaneously shortening the testing and training phases. In [13], researchers introduced a novel hybrid ID technique that combines anomaly and abuse detection models in a hierarchical decomposition framework. The C4.5 DT abuse detection model was initially developed for this research; it was then utilised to partition the entire training dataset into more manageable chunks. The one-class SVM was then used to build an anomaly detection model for each reconstructed region (1-class SVM). Using the behaviour-related information generated by malware execution on a host system, another study [14] proposes an anomaly-based detection technique to identify malware. This innovation aims to produce detection rules in abuse detection systems, and a work by [15] introduced Dendron as a novel way for GA-based evolving of DT classifiers. The proposed system provides rules that can be understood in any language, which benefits security administrators and lowers their risk. DT and improved fast heuristic clustering have been used to create a multi-level hybrid classification model, which has been presented [16]. This method has been demonstrated to be effective in detecting intrusions with a low false negative alert rate of 2.7% and a false-alarm rate of up to 9.1%, both of which are considered acceptable. When compared to the single-level method, this strategy excels in terms of ID performance and strikes a good balance between the rates of false negative and false positive alerts. This method's innovative feature incorporates both unsupervised (tree classifier design) and supervised learning (clustering analysis). A data mining-based architecture [17] consisting of two Bayesian networks and C5.0 structures has been proposed for use by the IDS. To maximize the strengths of both strategies, the system makes use of tree-augmented Nave Bayes and the boosting approach.

## 5. IDS BASED ON BAYESIAN CLASSIFIER

The foundation of the Bayesian classifier is the idea that a (natural) class should be able to predict the feature values of its members. As a result, many IDS experts have relied on Bayesian Classifier to sort network actions into "normal" and "abnormal" categories. In their research, the authors of [18] introduced a two-stage classifier that uses clustering to derive related subsets of system calls and models randomly. Together, supervised learning and clustering help identify suspicious patterns, while the incorporation of domain-level knowledge is achieved using suitable metrics. Numerous studies have worked to increase the Bayesian approach's accuracy in detecting R2L attacks and the rate at which the four most common types of intrusion may be identified. For example, data categorization powered by the Bayesian approach was achieved in [19]. With three characteristics (23, 24, and 31) and a threshold value of 0.6, the method showed improved performance for the R2L attack. An additional IDS-based work [20] proposed a BN classifier BMA for ID that uses the k-best BNs. In this analysis, we uncovered the DP algorithm that BMA employs to construct a BC from the global k-best structures. The research demonstrated that the BNMA classifier, which was built via a heuristic technique, has the superior predictive ability to the BN classifier and the Naive Bayes classifier. When compared to the other two classifiers trained with a larger dataset, the performance of the least-trained one is significantly better. The IDS field has been advised to use the Hidden Markov Model (HMM) framework [21]. The HMM applicator was trained and tested on the IDS dataset from the 1999 KDD Cup. From a total of 41 features in the dataset, only five were used in this study. Typical TCP link data was obtained from the KDD Cup 1999 data set, and the HMM was trained on this data; traffic was classified as either standard or attack. This demonstrated the HMM, with proper training and parameter estimates, as a potent tool for IDS development that can distinguish between benign and malicious traffic in real-time. Clustering-based two-stage classifier by[18] can generate similar system subsets and arbitrarily lengthy sequences as Markov chains.

## 6. IDS BASED ON OPTIMIZATION ALGORITHMS

Optimisation Various algorithms, including PSO, have found applications in IDS. When it comes to building models using rule generation, S-PSO has been shown to be the gold standard [22]. Using the KDD dataset, the suggested model's efficacy was evaluated, and the results showed that it was reliable and efficient. There were fewer false positives and improved overall detection accuracies in this model's rules. Additionally, [23] proposes a MapReduce-based ID system called IDS-MRCPSO. To control traffic on a massive network, this framework was created. This new technology was as effective as the popular MapReduce framework. Experiments conducted on a real-world intrusion dataset to evaluate the system speedup showed that the proposed IDSMRCPSO was effective even when the training dataset size was increased. [24] describes a different kind of hybrid IDS network. An intelligent dynamic swarm-based rough set (IDS-RS) is used to pick features, and simplified swarm optimisation with weighted local search (SSO-WLS) is used to classify intrusion data in this system. The created SSO-WLS relied on decision rules generation to boost anomaly detection efficiency. When tested on the KDD Cup 99 dataset, this technique proved to be reliable. With the WLS's help, we optimised the system's search process for SSO rule mining by adjusting the relative importance of the three fixed parameters. Essa [25] looked into using a combination of CFA and DT to choose features for ID. The suggested system's efficacy was tested on the KDD Cup 99 dataset. The CFA was used as a feature-selection tool before being tweaked and put to use with the DT classifier to quantify the output features.

## 7. IDS AIMS TO DECREASE THE DETECTION TIME

The time it takes to realise an attack has occurred is called the detection time. If an attack can be thwarted with greater efficiency, it will be less likely to succeed. Many different approaches have been tried and tested to accomplish this goal. Techniques like parallel processing and Big Data analysis tools are examples. After SVM parallelisation, reduced sample data was generated by principal component analysis (PCA), and the strategy was finally implemented on the Spark platform [26]. The experimental results validated the effectiveness of the proposed approach in speeding up the categorisation process with minimal loss of accuracy. Furthermore, a framework for swift and effective cybersecurity was described in [29]. The efficiency of the proposed framework was tested using Apache Spark and machine learning frameworks with KDD'99 and NSL-KDD datasets employing various classification and feature selection methods. By omitting these features from the used KDD'99 dataset, we could speed up the training and prediction processes while increasing accuracy. This paper demonstrates that a system's training and prediction times and accuracy are affected by the feature selection algorithm used, suggesting that future work could benefit from a more nuanced approach. The experimental findings for each feature selection method utilised in this study are listed in the table below.

TABLE I. PERFORMANCE EVALUATION USING THE CORRELATION-BASED FS METHOD ON THE KDD'99 DATASET

| Method | Accuracy | Training Time | Prediction Time |
|---|---|---|---|
| Logistic Regression | 91.56 | 289.105 | 12.909 |
| SVM | 78.84 | 479.124 | 10.085 |
| Naïve Bayes | 90.68 | 79.552 | 12.75 |
| Random Forest | 88.65 | 155.64 | 15.65 |
| GB Tree | 91.13 | 194.74 | 22.25 |

TABLE II. PERFORMANCE EVALUATION USING HYPOTHESIS TESTING-BASED FS METHOD ON THE KDD'99 DATASET

| Method | Accuracy | Training Time | Prediction Time |
|---|---|---|---|
| Logistic Regression | 91.64 | 320.256 | 15.686 |
| SVM | 92.13 | 530.45 | 19.02 |
| Naïve Bayes | 91.45 | 93.871 | 14.025 |
| Random Forest | 92.13 | 159.739 | 16.326 |
| GB Tree | 91.38 | 344.771 | 19.910 |

Finally, a comparison table (Table 3) was developed between most types of IDS and the methods used in the literature.

TABLE III. COMPARISON OF DIFFERENT IDS METHODS

| Auth. | Goal | Method | Accuracy % | | | |
|---|---|---|---|---|---|---|
| | | | Probe | U2R | Dos | R2L |
| **[4]** | Acc. | ANN | 98.79 | 96.51 | 99.93 | 99.54 |
| **[3]** | Acc. | GRNN | N.G | N.G | N.G | N.G |
| **[8]** | Acc. | SVM | 84.27 | 89.28 | 89.79 | 90.72 |
| **[17]** | A. | Bayesian | 100 | 93.75 | 100 | 99.64 |
| **[27]** | A. | Decision tree | 99.71 | 66.67 | 99.19 | 89.50 |
| **[28]** | A. | PSO | 97.8 | 88.7 | 99.9 | 70.1 |
| **[29]** | A. | Other approaches | 73.95 | 82.97 | 100 | 99.55 |
| **[30]** | Time | Parallel SVM on spark | 94.40 | 96.7 | 90.24 | 89.6 |

## 8. CONCLUSION

Intrusion detection systems can be constructed using various techniques, each with its own set of advantages and disadvantages depending on the goals it is meant to serve. In most cases, the outcomes and accuracy of hybrid methods (those that use elements from multiple approaches) are better than those of either individual approaches. Table 1 shows that the IDS that prioritises speedy detection falls short of the detection standards set by the IDS that prioritises precision. Future research is encouraged to strike a compromise between accuracy and the necessity to shorten the detection time by drawing on a variety of existing works; Tables 1 and 2 show that Naive Bayes, when running in parallel on Spark, is the best choice for intrusion prediction because it strikes a good balance between accuracy and speed. However, the results are also affected by the feature selection method, so the best results (accuracy and time) for intrusion detection would be achieved by using an appropriate and good feature selection method in conjunction with Naive Bayes running in parallel on Spark.

## Funding

None

## Conflicts Of Interest

The authors declare no conflicts of interest.

## Acknowledgement

The authors would like to thank the anonymous reviewers.

## References

[1] M. A. Mohammed, Z. H. Salih, N. Ţăpuş, and R. A. K. Hasan, "Security and accountability for sharing the data stored in the cloud," in *RoEduNet Conference: Networking in Education and Research, 2016 15th*, 2016, pp. 1-5: IEEE.

[2] H. Debar, M. Dacier, and A. Wespi, "A revised taxonomy for intrusion-detection systems," in *Annales des télécommunications*, 2000, vol. 55, no. 7-8, pp. 361-378: Springer.

[3] S. K. Gautam and H. Om, "Computational neural network regression model for Host based Intrusion Detection System," *Perspectives in Science,* vol. 8, pp. 93-95, 2016.

[4] I. Manzoor and N. Kumar, "A feature reduced intrusion detection system using ANN classifier," *Expert Systems with Applications,* vol. 88, pp. 249-257, 2017.

[5] R. A. R. Ashfaq, Y.-l. He, and D.-g. Chen, "Toward an efficient fuzziness based instance selection methodology for intrusion detection system," *International Journal of Machine Learning and Cybernetics,* vol. 8, no. 6, pp. 1767-1776, 2017.

[6] Z. Chiba, N. Abghour, K. Moussaid, and M. Rida, "A novel architecture combined with optimal parameters for back propagation neural networks applied to anomaly network intrusion detection," *Computers & Security,* 2018.

[7] H. Wang, J. Gu, and S. Wang, "An effective intrusion detection framework based on SVM with feature augmentation," *Knowledge-Based Systems,* vol. 136, pp. 130-139, 2017.

[8] M. S. M. Pozi, M. N. Sulaiman, N. Mustapha, and T. Perumal, "Improving anomalous rare attack detection rate for intrusion detection system using support vector machine and genetic programming," *Neural Processing Letters,* vol. 44, no. 2, pp. 279-290, 2016.

[9] P. Tao, Z. Sun, and Z. Suna, "An improved intrusion detection algorithm based on GA and SVM," *IEEE Access,* 2018.

[10] M. G. Raman, N. Somu, K. Kirthivasan, R. Liscano, and V. S. Sriram, "An efficient intrusion detection system based on hypergraph-Genetic algorithm for parameter optimization and feature selection in support vector machine," *Knowledge-Based Systems,* vol. 134, pp. 1-12, 2017.

[11] W. Shang, P. Zeng, M. Wan, L. Li, and P. An, "Intrusion detection algorithm based on OCSVM in industrial control system," *Security and Communication Networks,* vol. 9, no. 10, pp. 1040-1049, 2016.

[12] Y. Chen, Y. Li, X.-Q. Cheng, and L. Guo, "Building Efficient Intrusion Detection Model Based on Principal Component Analysis and C4. 5," in *Communication Technology, 2006. ICCT'06. International Conference on*, 2006, pp. 1-4: IEEE.

[13] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Systems with Applications,* vol. 41, no. 4, pp. 1690-1700, 2014.

[14] D. Moon, S. B. Pan, and I. Kim, "Host-based intrusion detection system for secure human-centric computing," *The Journal of Supercomputing,* vol. 72, no. 7, pp. 2520-2536, 2016.

[15] D. Papamartzivanos, F. G. Mármol, and G. Kambourakis, "Dendron: Genetic trees driven rule induction for network intrusion detection systems," *Future Generation Computer Systems,* vol. 79, pp. 558-574, 2018.

[16] L. Prema Rajeswari and A. Kannan, "An intrusion detection system based on multiple level hybrid classifier using enhanced C4. 5," *Communications and Networking Madras Institute of Technology. Chennai, India: IEEE,* pp. 75-79, 2008.

[17] F. Y. Nia and M. Khalili, "An efficient modeling algorithm for intrusion detection systems using C5. 0 and Bayesian Network structures," in *Knowledge-Based Engineering and Innovation (KBEI), 2015 2nd International Conference on*, 2015, pp. 1117-1123: IEEE.

[18] O. Koucham, T. Rachidi, and N. Assem, "Host intrusion detection using system call argument-based clustering combined with Bayesian classification," in *SAI Intelligent Systems Conference (IntelliSys), 2015*, 2015, pp. 1010-1016: IEEE.

[19] H. Altwaijry, "Bayesian based intrusion detection system," in *IAENG Transactions on Engineering Technologies*: Springer, 2013, pp. 29-44.

[20] L. Xiao, Y. Chen, and C. K. Chang, "Bayesian model averaging of bayesian network classifiers for intrusion detection," in *Computer Software and Applications Conference Workshops (COMPSACW), 2014 IEEE 38th International*, 2014, pp. 128-133: IEEE.

[21] N. Devarakonda, S. Pamidi, V. V. Kumari, and A. Govardhan, "Intrusion detection system using bayesian network and hidden markov model," *Procedia Technology,* vol. 4, pp. 506-514, 2012.

[22] Z. Yi and Z. Li-Jun, "A rule generation model using S-PSO for Misuse Intrusion Detection," in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, 2010, vol. 3, pp. V3-418-V3-423: IEEE.

[23] I. Aljarah and S. A. Ludwig, "Mapreduce intrusion detection system based on a particle swarm optimization clustering algorithm," in *Evolutionary Computation (CEC), 2013 IEEE Congress on*, 2013, pp. 955-962: IEEE.

[24] Y. Y. Chung and N. Wahid, "A hybrid network intrusion detection system using simplified swarm optimization (SSO)," *Applied Soft Computing,* vol. 12, no. 9, pp. 3014-3022, 2012.

[25] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems," *Expert Systems with Applications,* vol. 42, no. 5, pp. 2670-2679, 2015.

[26] C.-M. Ou, "Host-based intrusion detection systems adapted from agent-based artificial immune systems," *Neurocomputing,* vol. 88, pp. 78-86, 2012.

[27] A. H. Hamamoto, L. F. Carvalho, L. D. H. Sampaio, T. Abrão, and M. L. Proença Jr, "Network Anomaly Detection System using Genetic Algorithm and Fuzzy Logic," *Expert Systems with Applications,* vol. 92, pp. 390-402, 2018.

[28] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications,* vol. 60, pp. 19-31, 2016.

[29] G. P. Gupta and M. Kulariya, "A framework for fast and efficient cyber security network intrusion detection using apache spark," *Procedia Computer Science,* vol. 93, pp. 824-831, 2016.

[30] L. P. Rajeswari and A. Kannan, "An Intrusion Detection System based on multiple level hybrid classifier using enhanced C4. 5," in *Signal Processing, Communications and Networking, 2008. ICSCN'08. International Conference on*, 2008, pp. 75-79: IEEE.