Intrusion Detection using a Novel Hybrid Method Incorporating an Improved KNN

Hossein Shapoorifard
Department of Computer & Electrical Engineering
Shiraz branch, Islamic Azad University,
Shiraz, Iran

Pirooz Shamsinejad
Department of Computer Engineering & IT,
Shiraz University of Technology,
Shiraz, Iran

ABSTRACT

These days, with the tremendous growth of network-based service and shared information on networks, the risk of network attacks and intrusions increases too, therefore network security and protecting the network is getting more significance than before. Intrusion Detection System (IDS) is one of the solutions to detect attacks and anomalies in the network. The ever rising new intrusion or attack types causes difficulties for their detection, therefore Data mining techniques has been widely applied in network intrusion detection systems for extracting useful knowledge from large number of network data to detect intrusions. Many clustering and classification algorithms are used in IDS, therefore improving the functionality of these algorithms will improve IDS performance. This paper focuses on improving KNN classifier in existing intrusion detection task which combines K-MEANS clustering and KNN classification.

Keywords

Intrusion Detection System; Data Mining; Network Security; Clustering; IDS system; K-MEANS; K- nearest neighbor; K-farthest neighbor; CANN.

1. INTRODUCTION

With the tremendous expansion in computer network resources in recent years, a variety of network-based applications have been developed and published to provide services in different aspects such as social media services, banking services, government services, ecommerce services, etc. this application made Communication system to play a huge role in human’s daily life. Computer networks are widely used for business data processing, collaboration, education and learning and entertainment. This extensive use of computer networks made intruders to use different intrusion methods to access valuable network data. There are many devices and techniques which can be used to protect the network such as firewall, antiviruses, Intrusion detection Systems (IDS) and etc.

Generally intrusion detection systems are a type of security management systems for computer networks [1]. Gaining unauthorized access to files, network and any other serious security threat can be detected by IDS, generally IDS can detect any activity that breaks the security policy from various areas within computer and network environment [2]. IDS attempts to recognize and then notify the users activity as either normal or anomaly by comparing the network connection records to the known intrusion patterns and signatures obtained from the human experts. As traditional methods cannot keep with faster and more complex networks, we concentrate on data mining based intelligent decision technology to make faster and more effective decisions [3]. Data mining based intrusion detection commonly categorized into two main approaches: signature based detection and anomaly based detection. In signature based detection there is a database of predefined signatures or patterns which are taken from characteristic features that represents a specific attack. This method compares network traffic with those signatures/patterns to detect attack. on the other hand the main goal in anomaly based detection is to build a profile that represents network normal traffic, and then this method identifies network activity deviations from this profile to detect attacks [4][5]. Both of this approaches have their own advantages and disadvantages, Signature based detectors are very effective in detecting known attacks that are predefined in database, but the main drawback of signature based approach is its inability for discovering novel and unknown attacks. on the other hand, although the anomaly detection method can detect unknown and novel attacks, but those profiles that we mentioned before can sometimes be inaccurate which results into generation of false alarms, considering normal data as an attack, and of course Profiles should be updated constantly [6], [7].

There are many proposed hybrid approaches which are using data mining techniques to solve signature and anomaly based detection problems and to maximize their advantages. Methods that are based on integrating and combining different techniques are showing better results. In this paper our focus is on an already existing method named CANN which is using *k*-MEANS clustering along with *KNN* classifier [8].our goal is to improve *KNN* classifier performance to increase accuracy and detection rate and reduce false alarm rate of this method. in order to achieve this objective, we involved another effective factor in addition to nearest neighbor(KNN) to our classification process, that factor is farthest neighbor and we named this technique k farthest neighbor or *k*-FN. Our experimental results on NSL-KDD dataset shows that Involving farthest neighbor can increase accuracy and detection rate and reduce false alarm rate.

KDDCUP’99 is used worldwide for calculating the performance of various IDS [9] but statistical analysis on this data set, showed many issues which significantly affects the performance of evaluated IDS, therefore as we mentioned before we have used NSL-KDD dataset, NSL-KDD data set is a refined version of KDD99. It contains essential records of the complete KDD data set [10].

2. RELATED WORK

Machine Learning and Data Mining have provided powerful tools and techniques applicable for various range of applications upon variety of data types from electronic signals to internet documents to DNA sequences over the last two decades [11]–[21]. One of its influential application is for developing efficient intrusion detection systems (e.g. [2], [22], [5]-[8]). Om H et al. [23], offered a hybrid model that combines *k*-Means and two classifier methods: *k*-nearest neighbor and Naive Bayes. This model uses entropy based feature selection method for attribute selection. It applies *k*-

Means clustering algorithm for clustering purpose (used number of clusters five) which is followed by *k*-nearest neighbor (*KNN*) and Naïve Bayes classification algorithms for detecting intrusions. The model shows better approach than only *k*-Means. Author also used the KDD99 cup data set for performing their experiment.

Wang P and Wang J Q [24],discussed about data mining which is popularly known as an important way to mine useful information from large volumes of data which is noisy, fuzzy, and random. In this, present the whole techniques of the IDS along with data mining method in details. Author mainly discussed about three data mining based approaches: Classification, Association and Sequence rules. Also discussed the system architecture of the IDS.

Dewan et al. [25],proposed a learning algorithm for adaptive network intrusion detection using Naive Bayesian classifier and ID3 algorithm which performs good detections and keeps less false positives and also eliminates redundant attributes in addition to contradictory examples from training data set that make complex detection model. Author also addresses some difficulties of data mining such as handling continuous attribute, missing attribute values and reducing noise in training data. This model used Knowledge Discovery Data Mining (KDD) CUP 99 dataset for experiment.

Dhakar M, Tiwari A [26], presented an approach in perspective to enhance performance, the work presents a model for IDS. This improved model, named as REP (Reduced Error Pruning) based IDS Model gives output with greater accuracy along with the augmented number of properly classified instances. It uses the two algorithms of classification approaches namely, *k*2 (BayesNet) and REP (Decision Tree). Here REP provides an effective classification along with the pruning of tree with quick decision learning capability.

Amuthan Prabakar Muniyandi et al. [27] proposed an anomaly detection method using *k*-Means+C4.5 , a method to cascade k-means clustering and the C4.5 decision tree methods. This method achieves better performance in comparison to the *k*-Means, ID3, Naïve Bayes, *KNN*, and SVM.

Gisung Kim et al. [28], offers a new hybrid intrusion detection method hierarchically integrates a signature detection and anomaly detection in a decomposed structure. The signature detection model is built based on C4.5 decision tree algorithm and is used to decompose the normal training data into smaller subsets. The one-class SVM is used to create anomaly detection for the decomposed region.C4.5 decision tree does not form a cluster, which can degrade the profiling ability.

Jaiganesh et al [29] suggested a novel back propagation model for intrusion detection. This method makes training pair with a combination of input and equivalent target were generated and implemented into the network. Performance success can be measured by false alarm and detection rate. Detection rate was proven to be less than 80% for U2R, R2L, DoS and Probe attacks. However, the major issue of the method was found to be much inefficient to detect hidden attackers present in the system.

3. PROPOSED METHOD

As we mentioned before this method attempts to improve *KNN* classification process in a feature representation approach called CANN which is based on cluster center and nearest neighbor and now we involved farthest neighbor as another factor for more accrue classification. In this approach, two distances should be measured and summed:

- Distance between each data sample and each cluster center.
- Distance between data and its nearest neighbor in the same cluster.

As a conclusion it will induce a one-dimensional distance based feature which will be used to represent each data sample for intrusion detection by *k*-nearest neighbor (*KNN*) along with *k*-farthest neighbor (*k*-FN) classifier. As our experimental results based on the NSL-KDD dataset shows, in terms of classification accuracy, detection rates, and false alarms considering the farthest neighbor significantly improves detection rate, accuracy and reduces fails alarm rate. In the worst cases it performs similar to common *KNN* classification.

3.1 Producing single feature test and training dataset

As was said before, we used NSL-KDD dataset to examine our method performance. At first we need to preprocess and normalize our data from NSL-KDD. Linear transformation of the data, minimum maximum normalization I used for this step. This step is shown in “Figure 1”

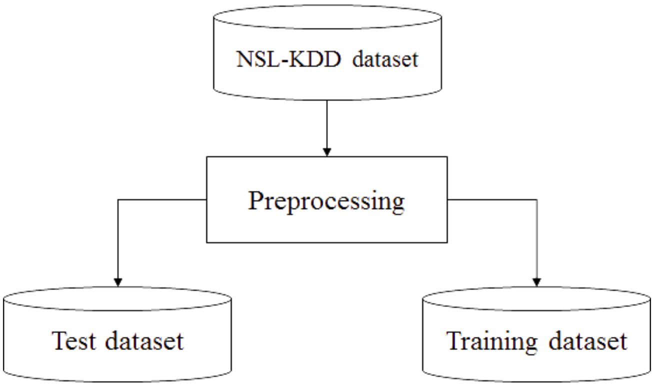

Figure 1. Data normalization

Next step is clustering our data *k*-MEANS algorithm is used for clustering our test and training data sets, our purpose is assigning the most similar data to a same cluster. We considered five as the number of our clusters *(k)*, the reason for this decision is that In the NSL-KDD data set there are four types of attacks beside the normal traffic

- U2R
- Probe
- Dos
- R2L

As a result, we are dealing with total of five types of network activity. Then we need to find the nearest neighbor to our data in its cluster and its distance to our data, we use *k* nearest neighbor to find it. Then we need to calculate the distance between our data and its cluster center and four other cluster centers, we use Euclid distance to calculate these distances. These distances are shown below in “Figure 2”

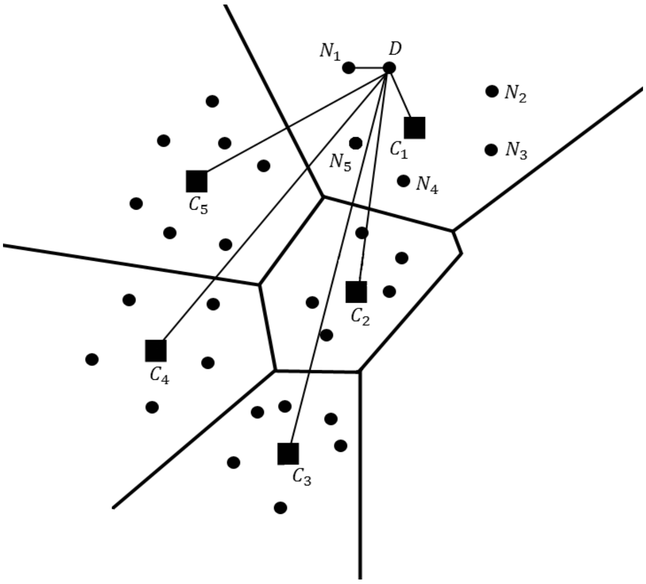

Figure 2. An example for distances between data and five cluster centers, data and its nearest neighbor

Considering *D* in "Figure 2" as our data, C_1 to C_5 as cluster centers and N_1 as the nearest neighbor to *D*. now the summation of these distances should be calculated. We can call the result of this summation D_T.

$$D_T = \overline{DC_1} + \overline{DC_2} + \overline{DC_3} + \overline{DC_4} + \overline{DC_5} + \overline{DN_1} \qquad (1)$$

Now D_T can be considered as a feature that can represent all other features, because all of them somehow affected D_T. We can repeat all of these steps for entire dataset to have single representative feature training and test datasets. You can see these steps in "Figure 3"

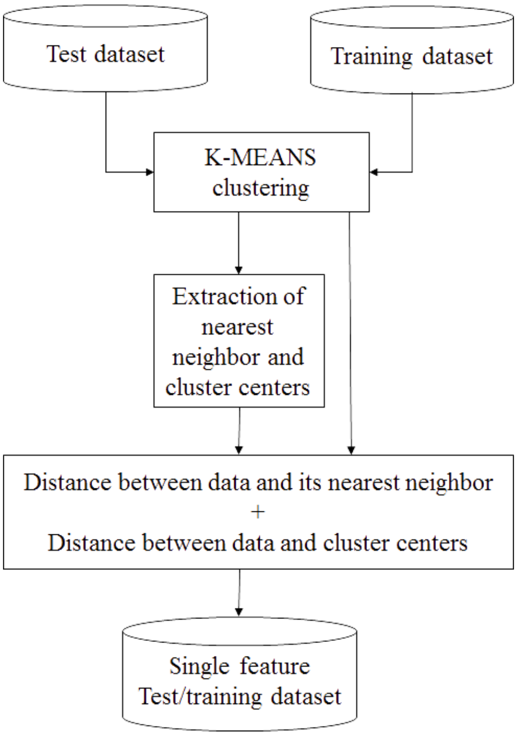
Test dataset
Training dataset
K-MEANS clustering
Extraction of nearest neighbor and cluster centers
Distance between data and its nearest neighbor
+
Distance between data and cluster centers
Single feature Test/training dataset

Figure 3. Building up single feature training & test datasets

3.2 Classification process

Now that our training and test datasets are transformed to single representative feature datasets, they are ready for classification based on nearest neighbor and farthest neighbor. Our proposed method is a distance based IDS, in distance based IDS we can assume that distance between normal and abnormal activity is big enough to make them distinguishable, so we can use distances to find out how much our test data is different or similar to our training data. As we mentioned before our proposed method uses both farthest and nearest neighbor to decide if our test data is an attack or not, therefore it is facing four possibilities.

- Nearest neighbor is normal, farthest neighbor is abnormal
- Nearest neighbor is abnormal, and farthest neighbor is normal
- Nearest and farthest neighbor are both normal
- Nearest and farthest neighbor are both abnormal

You can see an example of one of these possibilities in "Figure 4"

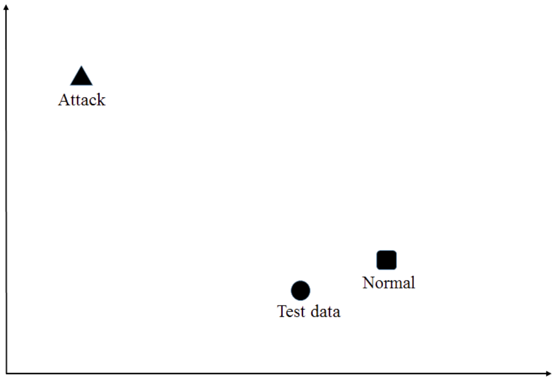

Figure 4. Farthest and nearest neighbor have different class labels.

Now we have to decide to classify our test data as an attack or as normal activity. As you can see in "Figure 4" our nearest neighbor is normal, and our farthest neighbor is an attack, in other words considering the mentioned assumption about distance based IDS, the most similar training data to our test data is normal and the most different training data from our test data is an attack, as a result of this explanation we can consider our test data as a normal activity. We can repeat this solution when our nearest neighbor is abnormal and the farthest one is normal, in this case our test data will be considered as an attack because in this case our most similar training data is an attack and our most different training data is normal. But we can't use this logic in two other remaining possibilities. In these cases when both nearest and farthest neighbors are normal or they are both abnormal we have to find another solution. We examined different solutions but the best solution was to classify our test data to a same class as its next (second) nearest neighbor. An example of this situation is shown in "Figure 5"

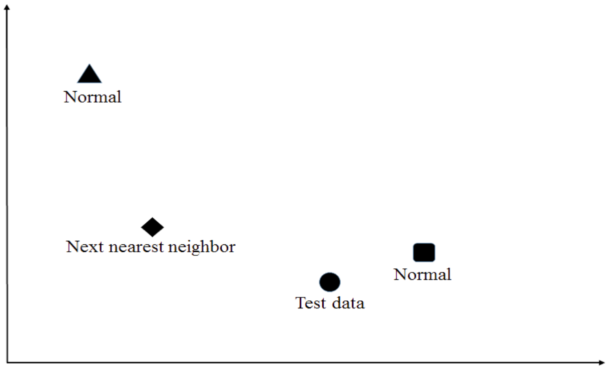

Figure 5. Farthest and nearest neighbor have similar class labels.

"Table 1" shows how our proposed approach classifies test data in each of mentioned possibilities, considering its farthest and nearest neighbor.

Table 1. Proposed approach decisions depending on different situations

Nearest and farthest neighbor class label		Class assigned to test data
Nearest neighbor	**Farthest neighbor**	
Normal	Attack	Normal
Attack	Normal	Attack
Normal	Normal	Depending on its next nearest neighbor
Attack	Attack	Depending on its next nearest neighbor

4. EXPERIMENTS AND RESULTS

Accuracy rate, detection rate and false alarm rate are three widely used IDS performance evaluation factors in intrusion detection studies. We also used these three factors to approximate our method's performance before and after improving classification process by involving farthest neighbor. Our experimental results shows that involving the farthest neighbor improves accuracy rate and detection rate and reduces fails alarm rate in most cases and in worst cases it performs similar. "Figure 6"shows CANN detection rate improvement after involving farthest neighbor in classification process after we repeated the experiment for ten times.

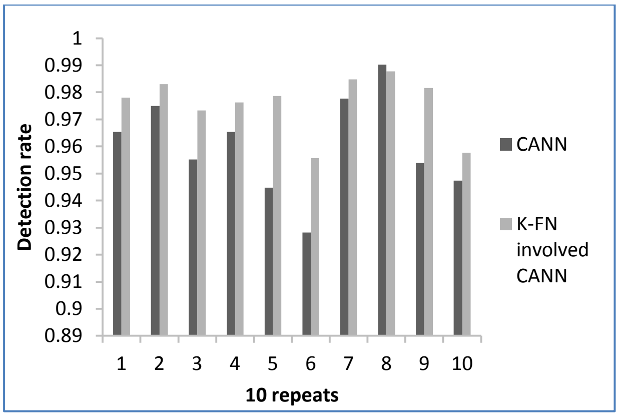

Figure 6. Detection rate, K-FN involved CANN vs CANN (higher is better)

"Figure 6" shows how much CANN performed better in terms of accuracy by involving farthest neighbor and next nearest neighbor.

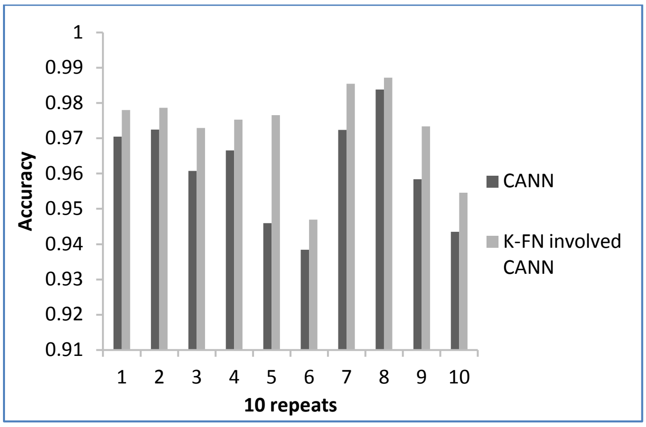

Figure 7. Accuracy rate, K-FN involved CANN vs CANN (higher is better)

The fails alarm rate also dropped significantly because by considering the farthest neighbor not only the most similar data affects the classification, but also the most different data is effective in this process. "Figure 8" shows significant fails alarm rate reduction.

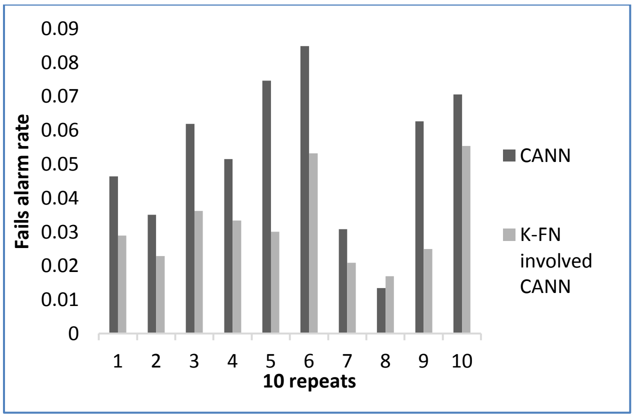

Figure 8. Fails alarm rate, K-FN involved CANN vs CANN (lower is better)

5. CONCLUSION

In this paper we have investigated some new techniques to improve classification performance in CANN intrusion detection approach and evaluated their performance on NSL-KDD dataset. We used the farthest neighbor (*k*-FN) along with nearest neighbor (*KNN*) for classifying our data, also we have used data's second nearest neighbor when both nearest and farthest neighbors had a same class label. Empirical results revealed that these new techniques improved or delivered equal performance in terms of accuracy, detection rate and reducing the fails alarm rate compare to direct *KNN* classification which was used in CANN, it was the prime concern of the proposed work. Future research work should pay closer attention to the data mining process. To deal with some of the general challenges in data mining, it might be best to develop special-purpose solutions that are tailored to intrusion detection.

6. REFERENCES

[1] J. Kim, P. J. Bentley, U. Aickelin, J. Greensmith, G. Tedesco, and J. Twycross, "Immune system approaches to intrusion detection - A review," *Natural Computing*, vol. 6, no. 4. pp. 413–466, 2007.

[2] M. Gupta, "Hybrid Intrusion Detection System: Technology and Development," *Int. J. Comput. Appl.*, vol. 115, no. 9, pp. 975–8887, 2015.

[3] M. Panda, A. Abraham, and M. R. Patra, "Procedia Engineering International Conference on Communication Technology and System Design 2011 Detection," vol. 0, no. 2011, 2012.

[4] E. Cloe and R. Krutz, *Network Security Bible*, vol. 1542, no. 9. 2015.

[5] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means," *Ain Shams Eng. J.*, vol. 4, no. 4, pp. 753–762, 2013.

[6] V. Golmah, "An Efficient Hybrid Intrusion Detection System based on C5. 0 and SVM.," *Int. J. Database Theory Appl.*, vol. 7, no. 2, pp. 59–70, 2014.

[7] J. Patel and K. Panchal, "Effective Intrusion Detection System using Data Mining Technique," vol. 2, no. 6, pp. 1869–1878, 2015.

[8] W. C. Lin, S. W. Ke, and C. F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-Based Syst.*, vol. 78, no. 1, pp. 13–21, 2015.

[9] M. K. Siddiqui and S. Naahid, "Analysis of KDD CUP 99 Dataset using Clustering based Data Mining," *Int. J. Database Theory Appl.*, vol. 6, no. 5, pp. 23–34, 2013.

[10] L. Dhanabal and S. P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 6, pp. 446–452, 2015.

[11] P. Mishra, E. S. Pilli, V. Varadharajan, and U. Tupakula, "Intrusion detection techniques in cloud environment: A survey," *J. Netw. Comput. Appl.*, vol. 77, pp. 18–47, 2017.

[12] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, 2017.

[13] A. Taheri and M. Shamsfard, "SBUEI: results for OAEI 2012," in *Proceedings of the 7th International Conference on Ontology Matching-Volume 946*, 2012, pp. 189–196.

[14] S. Olyaee, Z. Dashtban, and M. H. Dashtban, "Design and implementation of super-heterodyne nano-metrology circuits," *Front. Optoelectron.*, vol. 6, no. 3, pp. 318–326, 2013.

[15] M. H. Dashtban and P. Moradi, "A novel and robust approach for iris segmentation," *Int. J. Comput.*, 2011.

[16] A. Taheri and M. Shamsfard, "Instance coreference resolution in multi-ontology linked data resources," in *Joint International Semantic Technology Conference*, 2012, pp. 129–145.

[17] A. Taheri and M. Shamsfard, "Mapping farsnet to suggested upper merged ontology," in *Asia Information Retrieval Symposium*, 2011, pp. 604–613.

[18] M. Dashtban, M. Balafar, and P. Suravajhala, "Gene selection for tumor classification using a novel bio-inspired multi-objective approach," *Genomics*, 2017.

[19] M. H. Dashtban, Z. Dashtban, and H. Bevrani, "A novel approach for vehicle license plate localization and recognition," *Int. J. Comput. Appl.*, vol. 26, no. 11, 2011.

[20] S. Olyaee, Z. Dashtban, M. H. Dashtban, and A. Najibi, "Hybrid analytical-neural network approach for nonlinearity modeling in modified super-heterodyne nano-metrology system," in *Telecommunications (ConTEL), Proceedings of the 2011 11th International Conference on*, 2011, pp. 525–530.

[21] D. Faria *et al.*, "AML results for OAEI 2015.," in *OM*, 2015, pp. 116–123.

[22] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 16–24, 2013.

[23] H. Om and A. Kundu, "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system," in *2012 1st International Conference on Recent Advances in Information Technology, RAIT-2012*, 2012, pp. 131–136.

[24] W. Pu and W. Jun-qing, "Intrusion detection system with the data mining technologies," in *Communication {Software} and {Networks} ({ICCSN}), 2011 {IEEE} 3rd {International} {Conference} on*, 2011, pp. 490–492.

[25] D. M. Farid, N. Harbi, E. Bahri, M. Z. Rahman, and C. M. Rahman, "Attacks classification in adaptive intrusion detection using decision tree," *World Acad. Sci. Eng. Technol.*, vol. 63, no. 3, pp. 86–90, 2010.

[26] M. Dhakar and A. Tiwari, "A New Model for Intrusion Detection based on Reduced Error Pruning Technique," no. September, pp. 51–57, 2013.

[27] A. P. Muniyandi, R. Rajeswari, and R. Rajaram, "Network anomaly detection by cascading k-Means clustering and C4.5 decision tree algorithm," in *Procedia Engineering*, 2012, vol. 30, pp. 174–182.

[28] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Syst. Appl.*, vol. 41, no. 4 PART 2, pp. 1690–1700, 2014.

[29] V. Jaiganesh, P. Sumathi, and S. Mangayarkarasi, "An analysis of intrusion detection system using back propagation neural network," *2013 Int. Conf. Inf. Commun. Embed. Syst.*, pp. 232–236, 2013.