KCMC: A Hybrid Learning Approach for Network Intrusion Detection using K-means Clustering and Multiple Classifiers

S. Vahid Farrahi
M.Sc student
Shiraz University of Technology
Shiraz, Iran

Marzieh Ahmadzadeh
Assistant Professor
Shiraz University of Technology
Shiraz, Iran

ABSTRACT
A network Intrusion Detection System (IDS) is a security tool that acts as a defensive line. One of the most important challenges in network intrusion detection research area is designing an accurate intrusion detection system in terms of high detection rate, high accuracy and low false alarm rate. Hybrid learning approaches employ to deal with this challenge since, they have promising results in terms of detection rate, accuracy and false alarm rate. This paper, proposed a general structure of a hybrid learning approach. Then, the proposed approach has been implemented using K-means Clustering and Multiple Classifiers (KCMC). The data have been partitioned based on K-means clustering algorithm. Then, each partition classified using a distinct classifier. Naïve Bayes, Support Vector Machines and OneR classification algorithms have been used as the classifiers. The proposed hybrid approach has better results comparing to single classifiers in terms of detection rate, accuracy and false alarm rate. The detection rate of the proposed hybrid learning approach is 99.50%.

Keywords
Network Intrusion Detection, Hybrid Learning, Clustering, Multiple Classifiers, Network Security, Data Mining

1. INTRODUCTION
An Intrusion Detection System (IDS) is a security tool that acts as a defensive line against attackers [1, 2]. There are many vulnerabilities in a computer network. This means that, computer security is a so critical issue. An IDS is a vital tool that can help us to keep the network secure. The primary goal of an IDS is to automatically trigger an alarm when a suspicious activity occurs in the network.

There are two kinds of IDS: Signature-based (or misuse-based) IDS (SIDS) and Anomaly-based IDS (AIDS) [1-4]. A SIDS keeps the attack patterns in the signature database and tries to find a match between the attack patterns and a behavior. Since a SIDS has the patterns of well-known attacks in the signature database, this fashion of intrusion detection has high detection rate, accuracy and low false alarm rate. On the other hand, a SIDS is not able to detect novel attacks or unseen attacks.

An AIDS creates the normal behavior profiles and then tries to find the behaviors that deviate significantly from the normal profiles. The major disadvantage of an AIDS is low detection rate, accuracy and high false alarm rate in comparison of a SIDS. The primary advantage of an AIDS is the ability of detecting novel attacks.

Data mining techniques seek for the valuable knowledge and information in the databases. Data mining techniques have been widely used in many applications as well as network intrusion detection. The process of intrusion detection based on data mining techniques has four major steps [5]:

(1) Capturing data packets.

(2) Extracting features to describe the network data packets.

(3) Learning a model.

(4) Using the model for intrusion detection (Predict the normal and anomalous behaviors based on a signature-based model or differentiate between normal and anomalous behaviors based on an anomaly-based model).

There are a wide variety of data mining techniques that have been used in intrusion detection such as clustering and classification. The implementation of data mining based IDS have some difficulties such as low detection rate, accuracy and high false alarm rate [6].

Hybrid intrusion detection approaches have the best results in terms of detection rates and false alarm rate [7]. The inherent ability of hybrid approaches can be used in order to achieve higher detection rate and lower false alarm rate. It means that, hybrid learning approaches can overcome the inherent problems of implementing data mining based IDS. Different techniques such as combination of clustering and classification techniques can be used to form a hybrid learning approach [7].

This paper combined a clustering algorithm and multiple classifiers in order to form a hybrid learning approach. The proposed method is aimed at improving intrusion detection in terms of accuracy, detection rate and false alarm rate.

The remainder of the paper is organized as follows. Section 2 includes an overview of existing related works in hybrid approaches in network intrusion detection area. Section 3 describes the general structure of the proposed hybrid learning model. Section 4 explained the system details, which more details about the system implementation and simulation design are described in Section 5. Section 6 describes the simulation results. Finally, Section 7 concludes the work and describes the system potentialities and future works.

2. RELATED WORKS
As mentioned earlier, hybrid learning approaches have promising results in terms of detection rate, accuracy and false alarm rate in intrusion detection area. Anomaly learning approaches are able to detect novel attacks, the rate of false alarm using an anomaly approach is equally high. Therefore,

many researchers used hybrid approaches in network intrusion detection research area in order to achieve higher detection rate and lower false alarm rate.

K-means clustering algorithm is one of the most efficient data mining algorithms, which is proved a promising technique in intrusion detection [8]. Combining K-means algorithm with a classification algorithm such as OneR classification and Naïve Bayes classification can help us to improve intrusion detection. So, some of the previous hybrid methods combined K-means clustering algorithm and a classification algorithm to form a hybrid learning approach for network intrusion detection. This paper combined K-means Clustering algorithm and Multiple Classifiers (KCMC). Thus, the focus is on the previous works that used the combination of K-means algorithm and a classifier. In addition, other papers that proposed a hybrid algorithm have been reviewed in this section but the focus is on the combination of K-means and a classifier.

2.1 Combination of K-Means and a Classifier

In [9, 10] the authors proposed the hybrid learning approaches for network intrusion detection using K-means clustering and Naïve Bayes classification to improve intrusion detection. In fact, the proposed methods in [9, 10] are almost the same in fundamental solution. Both of the methods used the combination of K-means and Naïve Bayes to form a hybrid approach and focus on the improvement of the network intrusion detection in terms of accuracy, detection rate, and false alarm rate. The proposed methods differentiate between the normal and anomalous data in the clustering stage by clustering the data into normal and anomalous. In other words, the data were clustered based on their similarities and dissimilarities. This can help the classification algorithm to classify some of the misclassified data correctly in the subsequent stage. In the classification stage, the data classified based on the Naïve Bayes algorithm. The major differences between these papers are in the presenting of the results and data sets. In addition, reference [10] used a feature selection method in the pre-processing stage but the authors did not mention the feature selection method exactly.

In [11] the authors proposed a hybrid approach for network intrusion detection using K-means clustering and OneR classification. The fundamental solution is like the solution in [9, 10] but, their proposed method used OneR classification as the classifier. The proposed approach partitioned the instances into anomalous and normal clusters based on K-means clustering algorithm. It means that, after utilizing K-means, OneR classification algorithm used as the classifier. Consequently, some of the misclassified instances during the clustering stage may be correctly classified in the subsequent classification stage. The proposed method has better results than OneR classification algorithm while using as a single classifier.

In [12] the authors used weighted K-means and Naïve Bayes classification to form a hybrid intrusion detection approach. The other researchers leave out K-means algorithm without any pre-processing but their proposed method focused on the pre-processing stage and used weighted K-means instead of K-means in order to improve the intrusion detection. In weighted K-means, the weight of each feature shows the importance of the corresponding feature. C5.0 decision tree algorithm was used to obtain the weight of each feature in the pre-processing stage. In the clustering stage, weighted K-means separates the normal and anomalous the data. Finally, in the classification stage Naïve Bayes classifier was used as the classifier.

In [13] random forests and weighted K-means were used to form a hybrid learning approach. Based on the experimental results, the authors showed that the misuse detection based on the random forest has high detection rate and high false alarm rate. On the other hand, anomaly detection based on K-means algorithm has lower detection rate and lower false alarm rate, in comparison of misuse detection. Thus, they combined weighted K-means and random forests in order to achieve high detection rate and low false alarm rate. The proposed method is able to detect novel attacks.

2.2 Combination of Multiple Classifiers

In [14] the authors used a two-stage hybrid method for network intrusion detection. In the proposed method, Support Vector Machines (SVM) and Artificial Neural Network (ANN) combined to form a hybrid learner. In the first stage, the data classified into two classes, using SVM algorithm namely, normal and attacks. In the second stage of the proposed method, the attack data classify again and the attack types will be mentioned. So, the data that are known as the attacks in the first stage, classify again in the second stage. Experimental results show that combination of SVM and ANN is superior to SVM and ANN while using individually.

In [15] C4.5 decision tree and 1-class SVM have been used in order to propose a hybrid method. C4.5 decision tree algorithm was used for misuse detection. In addition, the normal data were separated into smaller data sets based on C4.5 algorithm. In other words, instead of one data set, multiple data sets were produced and multiple models were constructed for anomaly detection using 1-class SVM algorithm. 1-class SVM is an algorithm for anomaly detection. For each normal data set, a model was constructed based on 1-class SVM. This research used a misuse detection model to improve the anomaly detection models.

The hybrid learning approach that proposed in this paper combines K-means Clustering and Multiple Classifiers (KCMC). Three classifiers namely Naïve Bayes, Support Vector Machines (SVM), and OneR have been used in the proposed approach. In other words, KCMC creates a hybrid approach using K-means clustering algorithm and multiple classifiers instead of one single classifier. KCMC uses the potential ability of multiple classifiers for classifying the network traffic data more accurately in terms of higher detection rate, accuracy and lower false alarm rate. The innovation of our approach compared to previous works is the combination of K-means and multiple classifiers for data classification to form a hybrid intrusion detection approach instead of one classifier.

3. THE GENERAL STRUCTURE OF PROPOSED HYBRID APPROACH

The basic idea behind our proposed approach is that to employ the potential ability of a single classifier in data classification. For example, a single classifier may has the ability to classify normal behaviors more accurately than Denial of Service (DoS) attack type. On the other hand, another classifier may have the ability to classify DoS attack type more accurately than normal behaviors. As the clustering algorithms are able to partition the data based on their similarities, the data can be categorized based on their similarities using a clustering algorithm. It means that, the proposed approach uses a clustering algorithm for partitioning the similar data into the disjoint groups and uses the classifiers

for data classification. A distinct classifier can be chosen for classifying each bunch of data that are in a specific category.

More specifically, the proposed approach clusters the data in the training set in order to produce disjoint clusters. Then, the data set splits into m separate data sets based on the number of clusters (m is the number of clusters). Finally, the proposed approach constructs a different model for each cluster using a distinct classifier and the training set. Selection of the classifier depends on the potential ability of the algorithm in the separated data classification. In other words, a classifier is preferred, which has less training error while classifying the portioned data in the training set.

Generally, the proposed approach works as follows for network traffic classification:

Step1: A new traffic data receives to IDS.

Step2: The traffic data clusters based on the clustering algorithm and mentions its cluster assignment.

Step3: Based on the cluster assignment IDS decides that which model should classify the traffic data.

Step4: The traffic data classifies based on the model.

The general structure of proposed hybrid learning has been shown in the Fig. 1. As shown in Fig. 1, the proposed hybrid approach is able to use a distinct classifier for each bunch of the data. As it is a general structure, it does not mention any specific algorithm for the implementation of the proposed approach. More details about selection of the clustering algorithm and classifiers can be mentioned in the implementation phase of the approach.

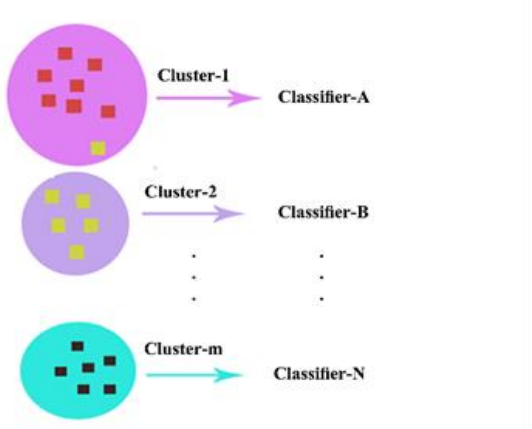

Fig. 1: General structure of proposed hybrid learning approach.

4. DETAILS ABOUT SYSTEM DESIGN

The previous section describes the general structure of the proposed hybrid learning approach. More details about the suggested clustering algorithm as well as the suggested classifiers presented in this section. It means that the described algorithms in this section have been used in the implementation of the proposed hybrid approach.

4.1 K-means Clustering Algorithm

K-means is a clustering algorithm that separates the data set into K disjoint clusters. K-means algorithm needs a distance metric for the computation of distances between the data points. Euclidean distance metric is the most common distance metric, which is used, in K-means clustering algorithm [16]. As mentioned earlier, it is proved that K-means algorithm has promising results in network intrusion detection [8]. The steps of K-means algorithm is shown in the following steps:

Step1 (initialization): Choose K randomly data points for initial cluster centers.

Step2 (assignment): Assign all data points to the nearest center based on the distance metric.

Step3 (updating): Update each cluster center by the mean of its cluster members.

Step4 (iteration): Repeat step 2 and 3 until reaching the stopping criteria or no more updating.

4.2 SVM, Naïve Bayes, and OneR classification algorithms

Classification is a data mining technique which operates on labeled data. Classification algorithm are able to predict a class label for each data point. Classification algorithms can be employed to predict intrusive behaviors in network intrusion detection. The main disadvantage of classification techniques is that they need labeled data for model construction.

SVM is one of the most successful classification algorithms in data mining area. SVM operates with statistical root and it is proved that has promising results in many practical applications [17]. SVM algorithm seeks for a hyperplane with maximal margin. In a linear separable case, there are many hyperplanes which might separate the data but the algorithm seeks for a hyperlane with maximal margin. Since, the hyperplane with maximal margin has better generalization error [17].

Naïve Bayes is another classification algorithm in data mining area that its construction is very simple. Naïve Bayes is based on a very strong independence assumption. A Naïve Bayes classifier estimates the class-conditional probability under the assumption that the attributes are conditionally independent [17]. It means that, the algorithm uses the relationship between independent variables and dependent variables to derive a conditional probability. Naïve Bayes classifier also have some inherent abilities such as robustness to isolated noise data points and robustness to irrelevant attributes [17].

OneR is a simple rule-based classification algorithm. The algorithm finds the most frequent class variable for each attribute value and make a rules for each attribute value using its most frequent class. Then, the algorithm calculates the error rate of each rule and finally the rule that has the smallest error rate will be picked.

5. SIMULATION DESIGN

To test the potential effectiveness of the proposed system in predicting network traffic data, the execution of K-means clustering algorithm and Naïve Bayes, SVM and OneR classification algorithms have been simulated. This section illustrates the simulation parameters and simulation details. In addition, the data set has been introduced in this section. As, the proposed approach employed K-means Clustering and Multiple Classifiers, the proposed method named KCMC.

5.1 Data set

The data for our experiments were produced by the 1998 DARPA intrusion detection evaluation program by MIT Lincoln Laboratory. The data set contains 4 attack types that is classified into four categories namely Denial of Service (DoS), Remote to User (R2L), User to Root (U2R) and

Probing. Each data point in the data set has 41 attributes, which show the characteristics of a network connection plus a class label. KCMC has been evaluated based on DARPA data set. 10% KDD cup 99 intrusion detection data set contains 494,021 data points [18].

The duplicate data points can cause the biased results of classifiers. Most of the data points are duplicate so, the duplicate data points have been removed from the data set. The remaining data points have been used for in our experiments. The selected data set contains 145,586 data points having 42 features. 72,784 data points have been used for the training phase of KCMC (training set) and 72,802 data points have been used in the testing (testing set) phase of KCMC.

5.2 KCMC Hybrid Learning Modeling Parameters

K-means clustering algorithm has been used for data clustering. Since, it has the ability of producing disjoint clusters and the algorithm is simple and efficient in intrusion detection area. KCMC partitioned the data set into four clusters (C1, C2, C3, C4) using K-means clustering algorithm. Since, U2R and R2l attack types have almost similar behaviors only one cluster considered for these attack types [11]. It means that the training set has been clustered into four disjoint clusters. Based on the clustering result each data point in the testing set has been clustered. Based on the cluster assignment of each test point the test set splitted into four disjoint data sets.

K-means clustering algorithm needs a distance metric for calculation of the similarities between the data. Euclidean distance metric was used as the distance metric in K-means clustering algorithm. Since the attributes of the data set are not in a specific range, an attribute may dominate the other attributes. To avoid the problem of attribute domination, the data were normalized based on the Min-Max method in the pre-processing phase.

As mentioned earlier, clustering algorithm has been used for categorizing the similar data. Since, K-means is an unsupervised algorithm the class label attribute has not used for the training and testing of K-means algorithm.

Naïve Bayes, SVM and OneR classification algorithms have been used in KCMC as the classifiers. SVM classification algorithm has been used for classifying the data which are in cluster C1. Also, OneR classification algorithm has been used in order to classifying the data which are belong to clusters C2 and C3 as well as Naïve Bayes algorithm for cluster C4.

The classifiers learned based on the training set. For example SVM classification algorithm has been learned using the data in the training set which are belong to cluster C1 and finally, the data which are belong to cluster C1 in the testing set have been classified using SVM. The same process repeated for the other three clusters but the classifiers have been changed. For data points, which belong to clusters C2 and C3, OneR algorithm has been used as the classier and Naïve Bayes algorithm has been used for classifying the data points, which belong to cluster C4.

5.3 Comparing Systems and Performance Evaluation

KCMC has been evaluted in terms of accuracy, Detection Rate (DR) and False Alarm Rate (FAR). In addition, this section compares KCMC with Naïve Bayes, SVM and OneR classifiers when uses as a single classifier. The same training set and testing set have been used for single classifiers and KCMC evaluation.

Performance evaluation metrics are defined as follows:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (1)$$

$$\text{Detection Rate (DR)} = (TP) / (TP+FP) \quad (2)$$

$$\text{False Alarm Rate (FAR)} = (FP) / (FP+TN) \quad (3)$$

Where

- True Positive (TP): Corresponds to the number of attacks that trigger an IDS to produces the alarms correctly.
- True Negative (TN): Corresponds to the number of normal activities that detected as normal by an IDS correctly.
- False Positive (FP): Corresponds to the number of normal activities that trigger an IDS to produces the alarms.
- False Negative: Corresponds to the number of attacks that an IDS did not detect.

6. RESLUTS

Table 1 represents the results across all category classes obtained from Naïve Bayes, SVM and OneR classification algorithms and the proposed approach. Naïve Bayes algorithm have the best result in term of accuracy in R2l and U2R attack types in comparison of the other two single classifiers. SVM algorithm has high accuracy in the predication of normal data and DoS attack type but is not able to predict any U2R attack. The accuracy of OneR algorithm in normal data is better than Naïve Bayes but OneR cannot predict any R2l and U2r attack types.

Table 1: Accuracy of Naïve Bayes, SVM, OneR and KCMC for different attack types.

Attack type	Naïve Bayes (%)	SVM (%)	OneR (%)	KCMC (%)
Normal	81.10	99.54	98.80	99.66
DoS	95.97	98.13	94.69	99.90
Probe	85.65	91.18	37.98	94.76
R2L	40.00	22.50	0.00	66.09
U2R	75.00	0.00	0.00	79.16

As shown in Table 1, KCMC performs better in all attack types in term of accuracy than all the other tree classifiers. KCMC is able to approximately predict all normal data and DoS attacks. It also performs better in the other attack types comparing the other three single classifiers.

The detection rate and false alarm rate of the single classifiers and proposed approach have been shown in Table 2. SVM performs well in detection rate in comparison of the other two classifiers. On the other hand, Naïve Bayes has the worst result in term of the detection rate. In addition, SVM has the lowest false alarm rate as a single classifier and Naïve Bayes has the highest false alarm rate. As a single classifier and considering false alarm rate and detection rate, SVM performs better than the other two classifiers. Although SVM performs

well in the terms of detection rate and false alarm rate, KCMC is superior to other tree classifiers. The detection rate of KCMC is 99.50%, which means that KCMC can predict approximately all of the attacks correctly.

Table 2: Detection rate and false alarm rete of Naïve Bayes, SVM, OneR and KCMC

Performance Metric	Naïve Bayes	SVM	OneR	KCMC
Detection Rate (%)	85.37	99.28	98.04	99.50
False alarm Rate (%)	18.98	0.45	1.19	0.33

The accuracy across all attack categories for single classifiers and KCMC has been compared in Fig. 2. KCMC is superior to the other three classifiers and its accuracy is 99.50%. Therefore, the proposed method has the potential ability to predict almost all of the normal data and attack data correctly.

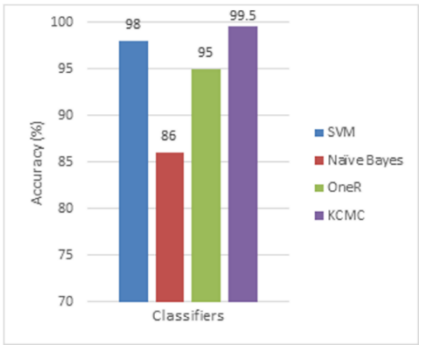

Fig. 2: Overall accuracy of KCMC and single classifiers

7. CONCLUSIONS AND FUTURE WORKS

In this study, a hybrid learning approach, which is named KCMC, has been presented. The model has investigated the combination of K-means clustering algorithm and three classifiers namely SVM, Naïve Bayes, and OneR. K-means clustering algorithm was used for producing the disjoint data sets. Multiple classifiers were used for data classification instead of one. The performance of the proposed approach has been evaluated based on the benchmark KDD Cup 99 intrusion detection data set.

Experimental results show that the proposed method is superior to the single classifiers in the terms of accuracy, detection rate, and false alarm rate. The innovation of our proposed method is the employment of multiple classifier instead of one classifier.

In this study, the proposed approach has been implemented using a clustering algorithm and three classifiers. The general structure is not limited to the preferred algorithms that, which have been used in the simulation of this paper. In other words, the clustering algorithm and classifiers can be changed based on the environment and implementation preferences.

Future work on this paper can focus on improving K-means clustering algorithm. K-means leave out algorithm without pre-processing. Pre-processing such as weighted K-means can improve the results of K-means algorithm. This can cause the data that are more similar will be assign to the same cluster. It is expected that the better data clustering will effect on better classification results.

In addition, K-Means algorithm can use for the detection of novel attacks. The proposed algorithm is not able to detect the novel attacks but its accuracy and detection rate are high. In the future works, the potential ability of K-means clustering algorithm can be used in order to detect the novel attacks.

8. REFERENCES

[1] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," computers & security, vol. 28, pp. 18-28, 2009.

[2] E. Biermann, E. Cloete, and L. M. Venter, "A comparison of intrusion detection systems," Computers & Security, vol. 20, pp. 676-683, 2001.

[3] A. Deepa and V. Kavitha, "A comprehensive survey on approaches to intrusion detection system," Procedia Engineering, vol. 38, pp. 2063-2069, 2012.

[4] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Computer Networks, vol. 51, pp. 3448-3470, 2007.

[5] W. Feng, Q. Zhang, G. Hu, and J. X. Huang, "Mining network data for intrusion detection through combining SVMs with ant colony networks," Future Generation Computer Systems, vol. 37, pp. 127-140, 2014.

[6] W. Lee, S. J. Stolfo, P. K. Chan, E. Eskin, W. Fan, M. Miller, et al., "Real time data mining-based intrusion detection," in DARPA Information Survivability Conference & Exposition II, 2001. DISCEX'01. Proceedings, 2001, pp. 89-100.

[7] K. Wankhade, S. Patka, and R. Thool, "An Overview of Intrusion Detection Based on Data Mining Techniques," in Communication Systems and Network Technologies (CSNT), 2013 International Conference on, 2013, pp. 626-629.

[8] M. Jianliang, S. Haikun, and B. Ling, "The application on intrusion detection based on k-means cluster algorithm," in Information Technology and Applications, 2009. IFITA'09. International Forum on, 2009, pp. 150-152.

[9] Z. Muda, W. Yassin, M. Sulaiman, and N. Udzir, "Intrusion detection based on K-Means clustering and Naïve Bayes classification," in Information Technology in Asia (CITA 11), 2011 7th International Conference on, 2011, pp. 1-6.

[10] S. K. Sharma, P. Pandey, S. K. Tiwari, and M. S. Sisodia, "An improved network intrusion detection technique based on k-means clustering via Naïve bayes classification," in Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on, 2012, pp. 417-422.

[11] Z. Muda, W. Yassin, M. N. Sulaiman, and N. Udzir, "Intrusion detection based on k-means clustering and OneR classification," in Information Assurance and Security (IAS), 2011 7th International Conference on, 2011, pp. 192-197.

[12] Y. Emami, M. Ahmadzadeh, M. Salehi, and S. Homayoun, "Efficient Intrusion Detection using Weighted K-means Clustering and Naïve Bayes Classification," Journal of Emerging Trends in Computing and Information Sciences, vol. 5, pp. 620-623, 2014.

[13] 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 Engineering Journal, vol. 4, pp. 753-762, 2013.

[14] J. Hussain, S. Lalmuanawma, and L. Chhakchhuak, "A Novel Network Intrusion Detection System Using Two-Stage Hybrid Classification Technique," IJCCER, vol. 3, pp. 16-27, 2015.

[15] 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, pp. 1690-1700, 2014.

[16] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern recognition letters, vol. 31, pp. 651-666, 2010.

[17] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to data mining vol. 1: Pearson Addison Wesley Boston, 2006.

[18] KDD cup99 intrusion detection data set. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup.data_10_percent.gz