A Cluster based Hybrid Framework for Network Intrusion Detection

Nusrat Mojumder
Ahsanullah University of Science and Technology

Md. Shahabub Alam
Ahsanullah University of Science and Technology

Mehtaz Afsana Borsha
Ahsanullah University of Science and Technology

Md. Mehedi Islam Khandaker
Ahsanullah University of Science and Technology

Syeda Shabanam Hasan
Ahsanullah University of Science and Technology

ABSTRACT
With the rise in storage and manipulation of sensitive data over networks and the colossal growth of network-based-services, security of network systems is being increasingly threatened. The necessity to create an efficient intrusion detection mechanism to detect cutting-edge cyber-attacks has become a daunting task for both the research community and the network industry. Various state-of-the-art methods have been employed in regards to solving this issues, Data-Mining being one of the most effective approaches. However, the generalization ability of individual data mining algorithms has limitations, and hence detecting complex attacks remains a daunting task. In such a context, this paper presents a novel hybrid technique based on the combination of both clustering and classification data mining approaches for developing an effective network intrusion detection system (NIDS) with increased accuracy and reduced false alarm rate. The models are trained and tested using the NSL-KDD intrusion detection dataset and information gain based feature reduction is used. In the result, a comparative study between different data mining classification methods after clustering is presented. Finally, it is experimentally prove that the proposed method is considerably more effective compared to some contemporary hybrid intelligence approaches.

General Terms
Network Security, Data Mining Algorithms, Machine Learning.

Keywords
Intrusion Detection, Network Security, Data Mining, Feature Selection, NSL-KDD, Information Gain, K-means clustering, Naïve Bayes, K-nearest-neighbor, Decision Tree, Support Vector Machine, Random Forest.

1. INTRODUCTION
Due to the recurrent growth and usage of internet technologies, network information has become more vulnerable to various cyber-attacks. It has become immensely important to make the network packets impervious to such attacks. Intrusion Detection has become crucial in protecting the systems against unauthorized and malicious activities.

The conception of network intrusion detection (NID) was proposed in 1980 by Anderson [1]. Intrusion detection is the procedure of supervising the events taking place in a computer system or network and scrutinizing them for indications of intrusions, characterized as attempts to compromise the confidentiality, integrity and availability of the system, or an attempt to evade the security measures of the system [2]. These intrusions are instigated by invaders gaining access to a system from the Internet, or by authorized users of the systems who attempt to gain additional benefits for which they are not permitted or by authorized users who misuse their privileges [3]. Intrusion detection systems (IDS's) dynamically monitor the activities happening on a network or information system, and conclude whether these activities are symptomatic of an attack or a legitimate use of the system [4]. IDS can be classified based on where the detection takes place: Host intrusion detection systems (HIDS) monitors the inbound and outbound packets on individual hosts or devices, in contrast Network intrusion detection systems (NIDS), are placed at strategic points within a network to observe and analyze the passing traffic for malicious activities.

Commonly, there are two types of approaches for intrusion detection system: Misuse or signature-based detection system identifies intrusion events that follow known patterns saved in a database of intrusions described as a sequence of actions or tasks that may be harmful. Anomaly based detection system, on the other hand, analyses network data and recognizes patterns of activities lying outside predefined normal events as possible intrusions or anomaly. Misuse detection ensures lower false alarm rates, but it is impossible to detect new attacks. Anomaly detection enables the detection of previously unknown attacks, it suffers from high false positive rates.

The purpose of this research is to build an anomaly based network intrusion detection system (NIDS) with high detection accuracy and low false alarm rate. Data mining techniques can be used to classify network connections into intrusion and normal data based on labeled training data in misuse detection [5], and to group similar network connections together in clusters according to a given measure in anomaly detection [6]. Good generalization ability to differentiate between normal and anomalous data and the ability accurately classify both known and unknown attacks dictates the effectiveness of such methods.

Even though classification is an effective method for defining groups or classes of objects, building classifier models by labeling a large number of records in the training dataset is costly. First partitioning the dataset into clusters based on similarity and then assigning labels to the relatively small number of groups is more effective and guarantees increased accuracy. Classification done after clustering helps improve results since the clustered results are more accurate with defining suitable decision boundaries and classification performed on these results will give better classified records and distinguish between class labels more effectively.

Intrusion detection systems are evaluated over datasets with an enormous amount of data with a number of various features. Some of the features may be may be redundant due

to high inter-relation with one of more of the other features while others could have poor prediction ability to the target patterns. These irrelevant features can affect detection accuracy and decrease detection rate. For achieving a better overall performance, any irrelevant and redundant features need to be discarded from the original feature space.

Using this knowledge, a novel approach for using the unsupervised clustering approach before supervised classification to improve classifier performance is presented in this paper. The clusters are formed primarily and are added to the dataset as an additional feature corresponding to the cluster number of each record. Classification is then performed on this dataset.

In this study, the fast K-Means clustering algorithm is used to separate and then label the data for the corresponding groups before applying classification. Five different classification algorithms: Naïve Bayes, K-Nearest-Neighbor, Decision Tree, Support Vector Machine and Random Forest are used as classification tools to classify normal and different types of attacks and compare results. The results are evaluated on the NSL-KDD intrusion detection dataset [7]. Pre-processing and normalization of dataset is performed and Information Gain technique is used to select relevant features.

The rest of the paper is organized as follows. Section 2 describes some related work for this research. Section 3 details the data mining algorithms used in this experiment. Section 4 provides the detailed analysis of the proposed algorithm. Section 5 describes the NSL-KDD dataset and experimental setup. Section 6 discusses the results followed by a conclusion in Section 7.

2. RELATED WORK

The concept of intrusion detection has existed for nearly two decades, but it garnered increasing amounts of attention with the rise of internet technology and attacks on network data. James Anderson [1] introduced the concept of Intrusion Detection in 1980, which discusses the model of observing the data over the local network by using predefined patterns. With the publication of this paper, the concept of intrusion and audit data came into discussion and officially laid the foundation of design and development of intrusion detection systems [8]. In the recent years, various data mining methods have been applied in the field of intrusion detection to determine whether the behavior of data is normal or an intrusion. The practice of using multiple algorithms for a hybrid approach has also become popular.

The authors of [9] present an IDS based on random forests and weighted k-means evaluated over the KDD'99 datasets. The proposed hybrid framework achieves 98% detection rates and 1.5% false alarm rates.

A hybrid data mining technique on a combination of k-means clustering and support vector machine classification used on the NSL-KDD dataset is presented in [10]. It has achieved 96.26 percentage in detection rate and 3.7 percentage as a false alarm rate.

The authors of [11] aim at proposing a hybrid of modified k-means with C4.5 intrusion detection system in a multi-agent system (MAS-IDS). KDD Cup 1999 dataset is used for evaluation with an accuracy of 91.13% and false alarm rate of 2.99%.

A hybrid approach is employed in [12] by clustering the data using the k-means algorithm and then classifying the data using the Adaptive-SVM algorithm. The experiment is carried out to evaluate the performance of proposed system is on NSL-KDD dataset achieving an accuracy of 98.47% and 0.53% false alarm rate.

The authors of [13] propose a hybrid learning approach through combination of k-means clustering and Naïve Bayes classification. The proposed approach clusters all data into the corresponding groups before applying a classifier. An experiment is carried out on KDD Cup '99 dataset show that the proposed approach performed better in term of accuracy, detection rate with reasonable false alarm rate.

The proposed method in [14] presents a k-means clustering algorithm based on particle swarm optimization (PSO-KM) that overcomes falling into local minima and has relatively good overall convergence. Experiments on data sets KDD'99 show the effectiveness of the proposed method and demonstrates higher detection rate and lower false detection rate.

An NIDS system based on SVM, Genetic Algorithm and Hierarchical Clustering is proposed in [15]. GA is used to eliminate the unimportant feature and BIRCH hierarchical clustering is used to provide optimal instances of the data set to the SVM. Evaluation on KDD'99 dataset show improved results.

Authors in [16] combine Decision trees (DT) and support vector machines (SVM) as a hierarchical hybrid intelligent system model (DT–SVM) and apply an ensemble approach in conjoining the base classifiers. Experimental results on the KDD Cup 1999 dataset exhibits that the ensemble approach uses the variances in misclassification and increases the overall performance by maximizing the computational proficiency and precision of detection for each class.

A new learning algorithm for adaptive intrusion detection using naïve Bayesian classifier and boosting that increases accuracy and reduces false alarms based on the KDD'99 dataset is introduced in [17].

The authors in [18] propose a hybrid method by linking X-Means clustering and Random Forest classification. Experimental results of XM-RF evaluated on the ISCX 2012 dataset show increased accuracy, detection and false alarm rates of 99.96%, 99.99%, and 0.2%, respectively compared with Random Forest and other hybrid approaches.

A hybrid algorithm of C5.0 and SVM is proposed by authors [19] to achieve better accuracy and detection rates. By combining the best results from the individual approaches, results reveal the hybrid C5.0–SVM algorithm delivers either equal performance or shows improvement while compared to the results of direct SVM approach for all the classes.

In [20] a combination of Naïve Bayesian (NB) and Support Vector Machine (SVM) as a hierarchical hybrid intelligent model is presented that maximizes the accuracy, which is the advantage of Naïve Bayes and reduces the false alarm rate which is the benefit of SVM.

A Classifier ensemble is designed in [34] using Radial Basis Function (RBF) and Support Vector Machine (SVM) as base classifiers. The experimental results demonstrate that the proposed ensemble methods provide significant improvement of accuracy compared to individual classifiers and also hybrid models display better results than standardized models for real and benchmark data sets of intrusion detection.

The authors in [35] propose an integrated machine learning algorithm using K-Means clustering and Naïve Bayes

Classifier called KMC+NBC. Experiments against the ISCX 2012 dataset show significant improvement in accuracy and detection rate up to 99% and 98.8%, respectively, while decreasing the false alarm to 2.2%.

3. DATA MINING ALGORITHMS

3.1 K-Means Clustering

K-means is one of the simplest unsupervised learning algorithms used for solving the clustering problem. The standard k-means clustering algorithm uses iterative refinement to produce a final result. The initial idea is to define centroids, one for each desired cluster and assign data points belonging to the given dataset to the nearest centroid. Next, new centroids are calculated using the previous assignments and data points are reassigned. This process is repeated iteratively where the centroids change their location step by step until the centroids do not move anymore and the algorithm converges.

Even though it is one of the most efficient clustering methods, the algorithm performs slowly and distance computations are high in practice for large datasets. In the paper, a triangle inequality based fast k-means clustering method is employed to overcome the limitations of standard k-means. Described in in Charles et al. [21], the triangle-inequality based k-means is an optimized version of the standard k-means method. The improved algorithm is based on the fact that most distance calculations in regular k-means are unnecessary. If a point is far away from a center, it is not necessary to calculate the exact distance between the point and the center in order to know that the point should not be assigned to this center. On the other hand, if a point is much closer to one center than to any other, calculating exact distances is not necessary to know that the point should be allocated to the first center. After each iteration, this process produces the same set of center locations as the standard k-means method, but it accelerates the process greatly.

3.2 Support Vector Machine

Support vector machine proposed by Vapnik in 1979 [22], is a potent statistical tool used to perform supervised classification. It is widely used for intrusion detection as classifier due to its good generalization ability.

The basic SVM is a binary classification problem that separates the data into two groups by means of a hyperplane using the support vectors. Support vectors are the data points that lie closest to the hyperplane and are most difficult to classify. The hyperplane that maximizes the distance to the nearest training data vector of any class achieves the best separation. In instances where SVM cannot separate two classes, it resolves this problem by mapping input data into high-dimensional feature spaces through nonlinear mapping using different kernel functions.

3.3 Decision Tree

Decision tree classifier is one of the predictive modeling approaches commonly used in data mining [23]. Tree models where the target variable can take a finite set of values are called classification trees. The objective of the classifier is to generate output predictions based on a set of inputs. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

The fundamental goal in a decision tree classifier is to come up with the optimal decision tree. Based on the values of attributes of the input, data is grouped together in Decision Trees. The decision tree classifier recursively picks the best attribute for splitting the data and growing the leaf nodes until the terminating condition is met. When all the data items in the current subset belong to the same class, the algorithm converges.

3.4 Random Forest

In 2001, Breiman [24] presented the concept of Random Forests which perform well in comparison to other data mining classification algorithms and can be implemented by bagging a set of decision trees. Random forests are a collaboration of tree predictors such that each tree depends on the values of an autonomously selected random sample. The distribution for all trees in the forest is equivalent. Each tree is constructed by taking a different section from the original data using a tree classification algorithm.

To classify a new object from an input vector, the input vector is put down each of the trees in the forest. Each tree provides a vote for a specific class which is the classification result of the tree. The classification with the maximum number of votes from all trees is chosen by the forest [24]. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [25].

3.5 Naïve Bayes

Naïve Bayes are a set of probabilistic classifiers based on utilizing Bayes' theorem using probability. Strong (naive) independence assumptions is used amongst the features that can predict a class, given a set of features using likelihood. Combining with statistical approaches, it plays an important role in intrusion detection, based on the effects of a probabilistic graphic model (PGM), in generating interdependencies between variables [26]. This graphical model does the learning job.

A probabilistic graphical model has nodes representing random variables (which may be discrete or continuous), and the edges representing conditional interdependency assumptions. Hence it represents a joint probability distribution. A conditional probability table is maintained for every variable, by the classifier, for which higher computational effort is required.

3.6 K-Nearest-Neighbor

The k-Nearest Neighbor algorithm (k-NN algorithm) is one of the most popular, straightforward and nonparametric data mining classification techniques. It is a simple algorithm that learns from the training examples and classifies new test examples based on a similarity measure in the feature space. Euclidean distance is used as a distance metric and similarity based search is used to discover the optimal hypothesis function. An object is categorized by the majority vote of its neighbors.

In the training phase, the input data points and class labels of the training data are stored. For classification, it computes the approximate distances between the new test data point and different points on the input vectors and then the test data is assigned to the most similar class amid its k nearest neighbors. Large prediction time is needed if the value of K is large [27]. The function is estimated locally and calculation is postponed until the classification is done.

4. PROPOSED METHODOLOGY

The proposed intrusion detection system is composed of various modules. Figure (1) shows a block diagram proposed intrusion detection system:

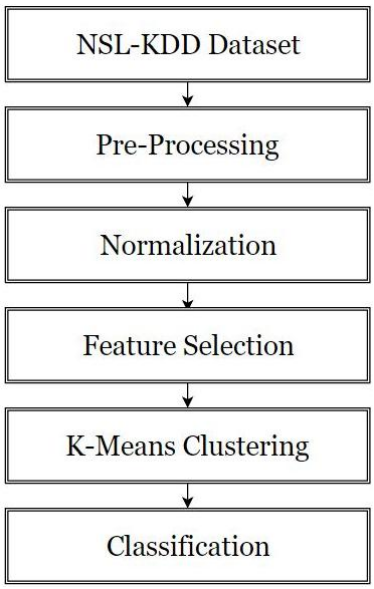

Fig 1: The Proposed Intrusion Detection System

4.1 Preprocessing

Pre- processing techniques help in removing the redundant or incomplete data, converting the non-numeric values to numeric values and transforming the data into a uniform format. The following preprocessing steps are performed on the NSL-KDD data set:

- Convert the non-numeric nominal features to numeric features. In NSL-KDD data set, all features of the data set take numeric or binary values apart from protocol type, service, and flag. These features are converted to numeric by replacing the unique values with their corresponding index numbers.
- Convert the class names to a numerical category by substituting them with numbers. Example: Replace 1 for Normal, 2 for DoS (Denial of service), 3 for probe, 4 for R2L (remote to-local) and 5 for U2R (user-to-root).

4.2 Normalization

The NSL-KDD data set contains both discrete and continuous attributes. There are different ranges of values for different attributes, making them incomparable and biased. In the paper, the features were normalized by using min-max normalization. By mapping all the different values for each feature within [0, 1] range, biased result was avoided.

4.3 Feature Selection

An effective feature selection scheme is vital in constructing a high performance IDS. Feature selection involves finding a subset of features to enhance prediction accuracy or decrease the size of the structure without considerably decreasing prediction accuracy of the classifier built using only the carefully chosen features [28]. Redundant features are usually closely correlated with one or more other features and omitting them does not damage classification accuracy. In fact, the accuracy may improve due to the removal of noise and measurement errors associated with the omitted features and resulting data reduction [29]. In the experiment, an Information Gain based feature reduction is applied.

For feature reduction, in a given set of feature vectors, the attributes most useful for differentiating between the classes is to be determined. Information gain gives the measure of how necessary a specific attribute of the feature vectors is. The basic concept of information gain is based on the decline in information entropy after a dataset is split on an individual attribute. It is denoted as [30]:

$$IG\ (T, a)\ =\ H(T)\ -\ H(T|a) \qquad (4.1)$$

The features are ranked based on their gain value, selecting the top 12 ranked attributes as the IG feature set.

4.4 Clustering

After feature reduction, clustering is used to group together the unlabeled data instances into clusters. Intrusions and normal instance are dissimilar in quality, so they do not fall into the same cluster. The clustering approach used for the experiment is triangle inequality based fast k-means clustering. K-means is fast, robust and in comparison to other clustering algorithms, shows remarkable supremacy. The runtime issues of k-means are overcome using the triangle inequality approach.

Clustering is used to generate a new set of features that corresponds to cluster numbers and incorporate into the dataset. This helps improve the classification result.

4.5 Classification

Finally, after clustering, classification algorithms are employed to build models that best fit the relationship between the attribute set and class label of the training data so that it can accurately predict the class labels of previously unknown test records. Different classification methods are chosen based on their good generalization ability, high accuracy rate in the field of intrusion detection and the diversity of their approaches in solving the problems. The results of these learning algorithms are compared to determine which hybrid clustering-classification model performs best based on the processed dataset.

5. EXPERIMENTAL SETUP

The thesis utilizes the NSL-KDD data set [31] which is a refined version of the original KDD'99 dataset [32]. It contains essential records of the complete KDD'99 data set but is devoid of the redundant and duplicated records, so the classifiers won't be biased towards more frequent records. There are a reasonable number of records in the train and test sets, thus it is affordable and cost effective to run the experiments on the entire dataset instead of randomly sampling a particular portion. Subsequently, evaluation results of different research works are consistent and can easily be compared against each other.

5.1 Dataset Description

5.1.1 Dataset Features

NSL-KDD data has 41 features and three features types: Numeric, Nominal, and Binary. Features 7, 12, 14, 15, 21, and 22 are binary, features 2, 3, and 4 are nominal and the remaining features are numeric. The nominal features are discrete and the numeric and binary features are continuous.

Among these 41 features [33]:

- 1-9 Basic features of each network connection vector
- 10-22 Content related features of each network connection vector
- 23-31 Time-related traffic features of each network connection vector
- 32-41 Host-based traffic features in a network connection vector

Table 1. NSL-KDD Dataset Features

No.	Feature Name	No.	Feature Name
1	duration	**22**	is_guest_login
2	protocol_type	**23**	count
3	service	**24**	srv_count
4	flag	**25**	serror_rate
5	src_bytes	**26**	srv_serror_rate
6	dst_bytes	**27**	rerror_rate
7	land	**28**	srv_rerror_rate
8	wrong_fragment	**29**	same_srv_rate
9	urgent	**30**	diff_srv_rate
10	hot	**31**	srv_diff_host_rate
11	num_failed_logins	**32**	dst_host_count
12	logged_in	**33**	dst_host_srv_count
13	num_compromised	**34**	dst_host_same_srv_rate
14	root_shell	**35**	dst_host_diff_srv_rate
15	su_attempted	**36**	dst_host_same_src_port_rate
16	num_root	**37**	dst_host_srv_diff_host_rate
17	num_file_creations	**38**	dst_host_serror_rate
18	num_shells	**39**	dst_host_srv_serror_rate
19	num_access_files	**40**	dst_host_rerror_rate
20	num_outbound_cmds	**41**	dst_host_srv_rerror_rate
21	is_host_login		

5.1.2 Dataset Classes

The 42nd attribute of the dataset contains information about the various class labels associated with the data. The NSL-KDD data includes 5 classes consisting normal and 4 types of attacks: Dos, Probe, R2L, and U2R. The four attack classes in the NSL-KDD dataset are:

- Denial-of-Service (DoS): The Denial-of-service (DoS) attack is an effort to make a network resources unobtainable to its intended users, to temporarily or indefinitely interject or suspend services of a host linked to the Internet. After gaining access to the network, the attacker can flood the network with traffic until overload triggers a shutdown, send invalid data to cause irregular termination or block traffic to renounce authorized users from accessing the network.
- Probing: In probing, the attackers survey the network for vulnerabilities and try to gather any potentially relevant information in the network. There are two kinds of Network Probe attacks: Host Sweep and Port Scan attacks. Host Sweep attacks are used to determine the hosts existing in the network, while port scan attacks are used to take advantage of the weaknesses of the host by detecting running services and exposed ports.
- User-to-Root (U2R): In User to Root attacks, the attacker starts out with gaining illegal access to a normal user account on the system and is able to exploit some vulnerability to achieve root access to the system. There are various types of User to Root attacks, most common being the buffer overflow attack. Buffer overflows transpire when a program replicas too much data into a static buffer without making sure that the data will fit.
- Remote-to-Local (R2L): A remote attack is a malicious intrusion that targets one or a network of computers. The remote attack does not influence the computer the attacker is using. Instead, the attacker will find susceptible points in a computer or network's security software to gain remote access to the machine or system. Remote attacks are generally carried out for illegally viewing or stealing data illegitimately or to introduce virus and other malicious events in a network.

5.1.3 Dataset Distribution

The entire NSL-KDD train and test dataset is used in the experiment. The data is distributed as following:

Training Data:

- 125973 records
- 53% are normal
- 47% are distributed among the different attack types

Test Data:

- 22544 test records
- 43% are normal
- 57% are distributed among the different attack types

5.1.4 Feature Selection

Information gain based feature selection technique is applied on the dataset, which is a ranking based feature selection technique that can be used to select features based on their rank. The feature with the highest information gain is placed on top and the rest are placed in descending order. The selected top features for this experiment are:

Table 2. Selected Features by Information Gain

No.	Feature Name	Gain
1	flag	0.9624
2	service	0.9100
3	diff_srv_rate	0.9020
4	same_srv_rate	0.8298
5	dst_host_diff_srv_rate	0.7848
6	dst_host_same_srv_rate	0.7241
7	dst_host_srv_count	0.6963
8	count	0.6834
9	dst_host_serror_rate	0.6777
10	serror_rate	0.6565
11	dst_host_srv_serror_rate	0.6177
12	srv_serror_rate	0.6048

6. RESULT ANALYSIS

The performance evaluation of the proposed IDS consists of two phases. First, the performance of the hybrid approaches in detecting anomaly is evaluated. Then the proposed experiment is compared to existing hybrid intelligent approaches.

6.1 Anomaly Detection Performance

In the first step, the individual algorithms are compared with all the hybrid approaches:

Table 3. Comparative Performance Analysis for Anomaly Detection

Hybrid Approach	Accuracy	False Alarm Rate	Precision	Recall
Support Vector Machine	71.56%	44.57%	61.19%	92.88%
FKM-SVM	99.38%	6.14%	99.32%	99.99%
Decision Tree	76.59%	34.74%	66.61%	91.59%
FKM-DT	99.73%	1.56%	99.83%	99.88%
Random Forest	75.24%	40.98%	63.87%	96.72%
FKM-RF	99.80%	2.00%	99.78%	1.00%
Naïve Bayes	74.67%	39.18%	64.24%	93.00%
FKM-NB	99.75%	2.54%	99.72%	1.00%
K-Nearest-Neighbor	78.70%	35.13%	67.63%	96.97%
FKM-KNN	98.95%	10.54%	98.85%	1.00%

Here, a relative measure of the performance of the different individual and hybrid classifiers in detecting anomalies is given. The experiments show that the hybrid methods significantly improve the performance of IDS.

Each hybrid classifier outperforms the individual approaches in terms of accuracy and there is a substantial decline in false positive rates. Thus, the goal of the experiment is achieved.

6.2 Performance Comparison with Existing Hybrid Intelligence Approach

Lastly, experimental result are compared with existing hybrid intrusion detection approaches to prove the authenticity of the experiment.

Table 4. Comparative Performance Analysis with Existing Methods

No.	Feature Name	
Hybrid Approach	**Dataset**	**Accuracy**
RBF-SVM [34]	NSL-KDD	98.46%
KM-SVM [10]	NSL-KDD	95.76%
KM-Adaptive SVM [12]	NSL-KDD	98.47%
KMC+NBC [35]	ISCX 2012	99.00%
FKM-SVM	NSL-KDD	99.38%
FKM-DT	NSL-KDD	99.73%
FKM-RF	NSL-KDD	99.80%
FKM-NB	NSL-KDD	99.75%
FKM-KNN	NSL-KDD	98.95%

The comparative analysis displays the presented methods outperform previously existing intrusion detection approaches and provide improved intrusion detection accuracy.

7. CONCLUSION

In this paper, the proposed experiment is a hybrid IDS that makes use of K-Means Clustering to include additional cluster information in the feature dataset and uses different classifiers for evaluation on the NSL-KDD dataset. It is empirically shown that, as a whole, the methods clearly outperform primary algorithms in terms of accuracy. Also, the result shows significant reduction of false positive rates.

In future, the proposed method will be tested on real world datasets to measure the practical applications of the experiment. The next step is coming up with a better system by combining the most successful algorithms found in this paper into a hybrid structure. Future aspirations include implementing intrusion prevention capabilities along with intrusion detection and testing on real life network traffic to advance the work further.

8. REFERENCES

[1] Anderson, J. (1980). Computer security threat monitoring and surveillance. Technical Report, James P. Anderson Co., Fort Washington, PA.

[2] Bace, R., & Mell, P. (2001). NIST special publication on intrusion detection systems. Booz-Allen and Hamilton Inc Mclean Va.

[3] Akbar, S., Rao, K. N., & Chandulal, J. A. (2010). Intrusion detection system methodologies based on data analysis. International Journal of Computer Applications, 5(2), 10-20.

[4] Debar, H., Dacier, M., & Wespi, A. (1999). Towards a taxonomy of intrusion-detection systems. Computer Networks, 31(8), 805-822.

[5] Zhang, J., Zulkernine, M., & Haque, A. (2008). Random-forests-based network intrusion detection systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(5), 649-659.

[6] Huang, J. Z., Xu, J., Ng, M., & Ye, Y. (2008). Weighting method for feature selection in k-means. Computational Methods of feature selection, 193-209.

[7] NSL-KDD dataset http://nsl.cs.unb.ca/NSL-KDD/ Last Visited: May 2016

[8] Ashoor, A. S., & Gore, S. (2011). Importance of intrusion detection system (IDS). International Journal of Scientific and Engineering Research, 2(1), 1-4.

[9] Elbasiony, R. M., Sallam, E. A., Eltobely, T. E., & Fahmy, M. M. (2013). A hybrid network intrusion detection framework based on random forests and weighted k-means. Ain Shams Engineering Journal, 4(4), 753-762.

[10] Md Tahir, H., Hasan, W., Md Said, A., Zakaria, N. H., Katuk, N., Kabir, N. F., Md Omar, H., Ghazali, O. & Yahya, N. I. (2015). Hybrid machine learning technique for intrusion detection system. 5th International Conference on Computing and Informatics (ICOCI).

[11] Laftah Al-Yaseen, W., Ali Othman, Z., & Ahmad Nazri, M. Z. (2015). Hybrid Modified-Means with C4. 5 for Intrusion Detection Systems in Multiagent Systems. The Scientific World Journal, 2015.

[12] Chahal, J. K., & Kaur, A. (2016). A Hybrid Approach based on Classification and Clustering for Intrusion Detection System. IJMSC-International Journal of Mathematical Sciences and Computing (IJMSC), 2(4), 34.

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

[14] Li, Z., Li, Y., & Xu, L. (2011, September). Anomaly intrusion detection method based on k-means clustering algorithm with particle swarm optimization. In Information Technology, Computer Engineering and Management Sciences (ICM), 2011 International Conference on (Vol. 2, pp. 157-161). IEEE.

[15] Bisen, M., & Dubey, A. (2015). An Intrusion Detection System Based On Support Vector Machine Using Hierarchical Clustering and Genetic Algorithm. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), 3(1).

[16] Peddabachigari, S., Abraham, A., Grosan, C., & Thomas, J. (2007). Modeling intrusion detection system using hybrid intelligent systems. Journal of network and computer applications, 30(1), 114-132.

[17] Farid, D. M., Harbi, N., Bahri, E., Rahman, M. Z., & Rahman, C. M. (2010). Attacks classification in adaptive intrusion detection using decision tree. World Academy of Science, Engineering and Technology, 63, 86-90.

[18] Juma, S., MUDA, Z., Mohamed, M. A., & YASSIN, W. (2015). Machine Learning Techniques for Intrusion Detection System: A Review. Journal of Theoretical & Applied Information Technology, 72(3).

[19] Golmah, V. (2014). An efficient hybrid intrusion detection system based on C5. 0 and SVM. International Journal of Database Theory and Application, 7(2), 59-70.

[20] Sagale, A. D., & Kale, S. G. (2014). Combining Naive Bayesian and support vector machine for intrusion detection system. IJCAT International Journal of Computing and Technology, 1(3).

[21] Elkan, C. (2003). Using the triangle inequality to accelerate k-means. In Proceedings of the 20th International Conference on Machine Learning (ICML-03) (pp. 147-153).

[22] Vapnik, V. N., & Kotz, S. (1982). Estimation of dependences based on empirical data (Vol. 40). New York: Springer-Verlag.

[23] Rokach, L., & Maimon, O. (2014). Data mining with decision trees: theory and applications. World scientific.

[24] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[25] Ho, T. K. (1995, August). Random decision forests. In Document Analysis and Recognition, 1995. Proceedings of the Third International Conference on (Vol. 1, pp. 278-282). IEEE.

[26] Subaira, A. S., & Anitha, P. (2014, January). 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 (pp. 274-280). IEEE.

[27] Tribak, H., Delgado-Marquez, B. L., Rojas, P., Valenzuela, O., Pomares, H., & Rojas, I. (2012, May). Statistical analysis of different artificial intelligent techniques applied to Intrusion Detection System. In Multimedia Computing and Systems (ICMCS), 2012 International Conference on (pp. 434-440). IEEE.

[28] Baig, Z. A., Shaheen, A. S., & AbdelAal, R. (2011). One-dependence estimators for accurate detection of anomalous network traffic. International Journal for Information Security Research (IJISR), 1(4), 202-210.

[29] Koller, D., & Sahami, M. (1996). Toward optimal feature selection. Stanford InfoLab.

[30] Information gain in decision trees https://en.wikipedia.org/wiki/Information_gain_in_decision_trees Last Visited: November 2016

[31] NSL-KDD dataset http://nsl.cs.unb.ca/NSL-KDD/ Last Visited: May 2016

[32] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on (pp. 1-6). IEEE.

[33] Dhanabal, L., & Shantharajah, S. P. (2015). A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering, 4(6), 446-452.

[34] Govindarajan, M. (2014). Hybrid intrusion detection using ensemble of classification methods. International Journal of Computer Network and Information Security, 6(2), 45.

[35] Yassin, W., Udzir, N. I., Muda, Z., & Sulaiman, M. N. (2013, August). Anomaly-based intrusion detection through k-means clustering and naives bayes classification. In Proc. 4th Int. Conf. Comput. Informatics, ICOCI (No. 49, pp. 298-303).