A HYBRID APPROACH FOR INTRUSION DETECTION SYSTEM USING K-MEANS AND RANDOM FOREST

Pavan Kumar Suda,
Student, Computer Science and Engineering, ST.MARY'S GROUP OF INSTITUTIONS,Chebrolu,Guntur(DT)
pavankumarsuda@yahoo.com

Dr.G.Jaideep
Associate Professor, Computer Science and Engineering, ST.MARY'S GROUP OF INSTITUTIONS,Chebrolu,Guntur(DT)
, gera.jaideep@gmail.com

Abstract

Intrusion detection systems play important role in real world applications. Every organization or government that uses any sort of networking and information systems need protection from various kinds of intrusions. Many existing intrusion detection systems provide very highly verbose output and it is not easier for administrators to identify the issues immediately. With the Artificial Intelligence (AI) techniques with underlying Machine Learning (ML) algorithms, there is scope of developing IDS based on AI. In this project, a hybrid IDS is developed using machine learning approaches. It combines Random Forest classification and K-Means clustering. This will use both misuse detection and anomaly detection for improving performance of the IDS. These algorithms are evaluated for the four categories of attacks based on precision, recall, F1-score, false-alarm-rate, and detection-rate. The proposed IDS is evaluated with NSL-KDD dataset which is highly optimized for intrusion detection research. The results of experiments showed that the hybrid IDS performs well in terms of detection rate and other metrics.

***Keywords** –Intrusion Detection System (IDS), Random Forest (RF), K-Means clustering, machine learning*

1. INTRODUCTION

Machine learning is widely used for discovering business intelligence from data of different domains. In fact, organizations solve many problems with machine learning techniques that are part of AI. There are different kinds of techniques as shown in Figure 1. Broadly, they can be classified into supervised learning (training is needed), unsupervised learning (no explicit training), semi-supervised learning (have qualities of supervised and non-supervised) and reinforcement learning. In this project supervised learning and unsupervised learning models are used to realize a hybrid IDS. The IDS is made up of K-Means clustering and Random Forest (RF) classification algorithm. In most of the applications, classification accuracy can be improved with certain pre-processing techniques and clustering as well.

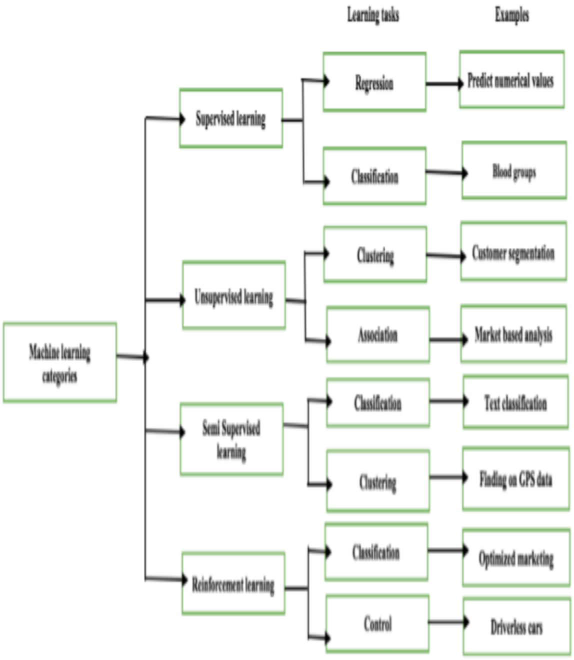

Figure 1: Various categories of ML techniques

As presented in Figure 1, there are many examples of applications for which different learning models are used. In this project K-Means clustering, RF classification and attribute ratio based feature selection are used. A hybrid IDE is realized using aforementioned techniques. The contributions in this project include the usage of K-Means and RF for making an IDE that is supported by attribute ratio which is a good feature selection method. If feature selection is not used, the IDS may show poor performance. Our contributions are as follows.

1. An IDS is built with K-Means clustering with attribute ratio-based feature selection.
2. An IDS is built with RF classification with attribute ratio-based feature selection.
3. A Hybrid IDS is realized by combining K-Means and RF for better results with different datasets.

The remainder of the report is structured as follows. Chapter 2 reviews literature. Chapter 3 provides details of the K-Means and RF. Section 4 presents the proposed system. Section 5 provides performance metrics. Section 6 covers experimental setup and dataset details. Section 7 presents experimental results. Section 8 concludes the paper along with directions for future work.

2. RELATED WORK

Different algorithms in ML are used for security analysis. Many IDS strived to reduce false alarm rate as explored in [1]. In [2] an IDS is realized with clustering technique and a classifier known as KNN. K-Means is used for intrusion detection in [3] which is widely used as a clustering technique in the data mining domain. The combination of K-Means and SVM are used in [4] for developing an IDS with extreme learning procedure. RBF kernel function and K-Means are used in [5] for developing an IDS with feature selection as well. The C4.5 and K-Means algorithms are studied for IDS in [6] while a Network IDS (NIDS) is developed in [7] based on weighted k-Means and RF.

Apache Spark is used in [8] for an effective IDS towards NIDS for cyber security. There are many classification techniques used for IDS as reviewed in [9]. A supervised tree based mechanism is employed in [10] while RF is used for NIDS in [11]. Botnet based attacks are explored and an IDS is made to detect such attacks using RF in [12]. AdaBoost and Artificial Bee Colony combination is employed in [13] for NIDS. Various data mining approaches are used in [14] for realizing an IDS. A hybrid mode for IDS with feature selection is considered in [15] while the combination of SVM and GA (Genetic Algorithm) are used for IDS in [16]. Intrusion detection for Wireless Sensor Network (WSN) [17], the combination of anomaly detection and misuse detection are explored in [18] while a hybrid IDS for power system is made in [19]. From the literature, it is understood that there is need for hybrid IDS with feature selection for higher level of accuracy and least false alarm rate.

3. PROPOSED HYBRID APPROACH

The methodology used for building an intrusion detection system based on machine learning algorithms is described here. Figure 2 shows outline of the IDS. It takes data as input and performs intrusion detection process. It can distinguish between the malicious and benign instances. The dataset used for the experiments is NSL-KDD taken from [20].

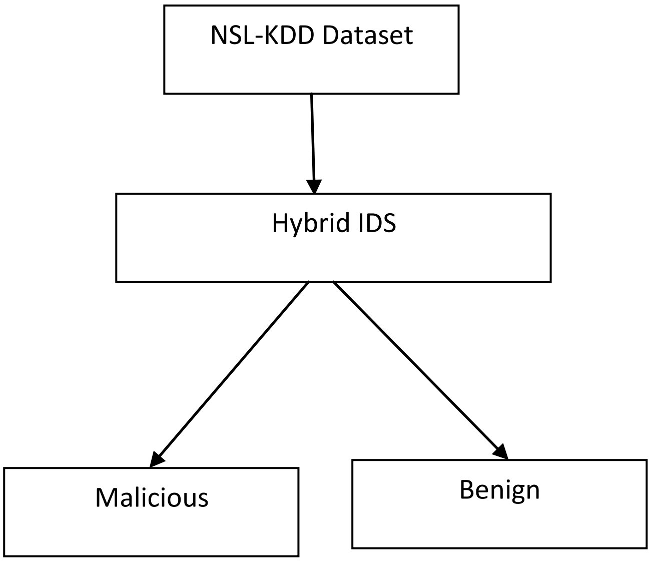

Figure 2: Outline of IDS

The intrusion detection system is provided with more details in Figure 3. It consists of many functions like pre-processing of data to have training and testing sets, feature extraction, feature selection and intrusion classification. Feature extraction is the process of obtaining all the features from the NSL-KDD dataset. Once all the features are extracted, there is need for feature selection. Feature selection is the process in which some features that have higher weight in contributing towards detecting class labels. There are many feature selection methods such as filter and wrapper methods. One filter method is employed in this project. It is known as attribute ratio method. According to this method the extracted features are subjected to an iterative process in which each attribute is verified to know whether it is having any contribution towards detecting class label. Once the features are identified, the selected features are provided to the classifier for training. In other words, the training data with chosen features are provided to classifier. Here the classifier used is RF algorithm. It takes the selected features as part of training set and learns from the data. As the training data contains class labels, the algorithm learns from it and makes a model. This model is known as knowledge model that is used for prediction of class labels for unlabelled data. Once training is completed, in the testing phase, the RF classifier will be able to show the class labels for unlabelled instances. Thus the proposed IDS is able to identify intrusions.

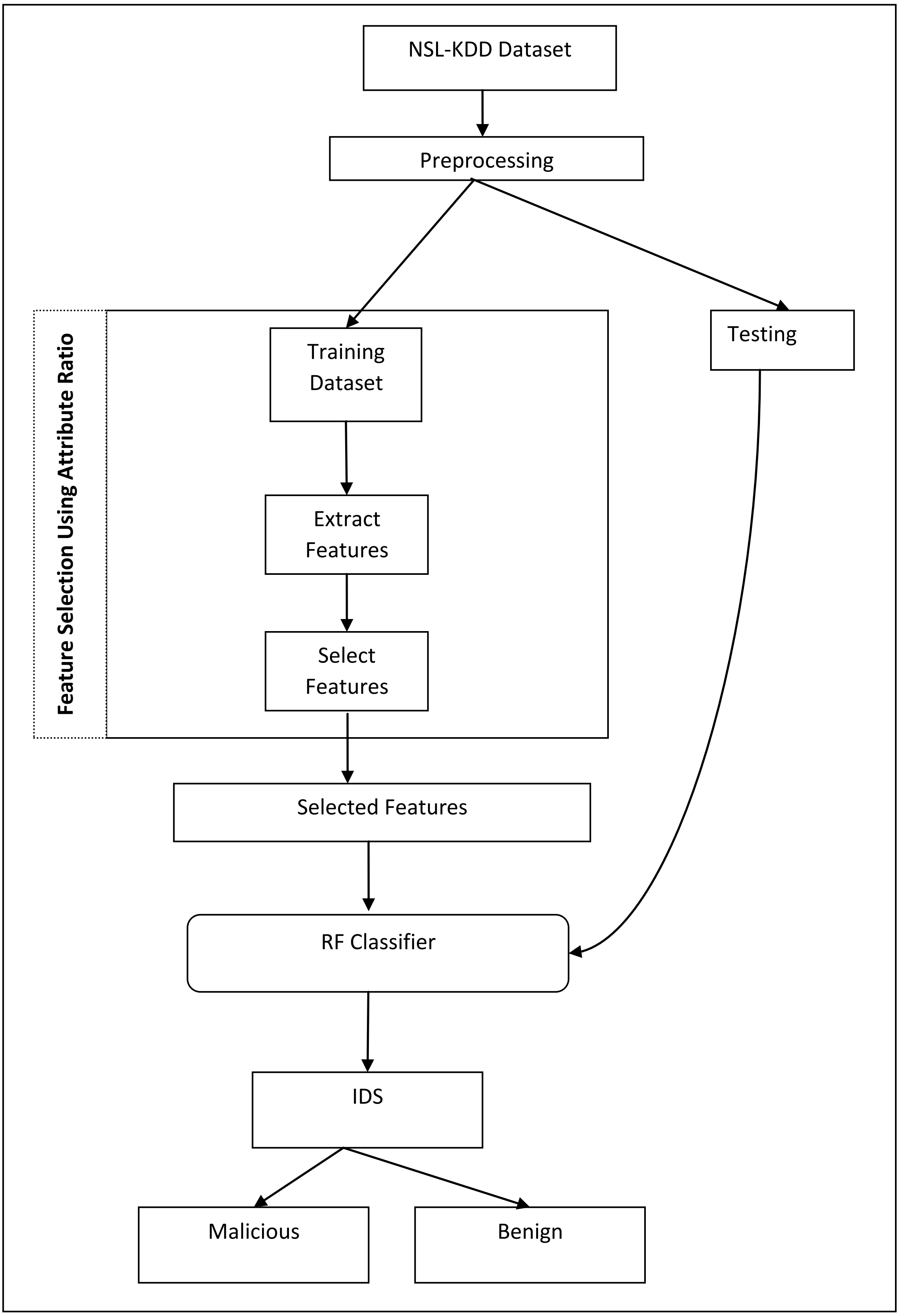

Figure 3: IDS with RF classifier

The RF classifier is used as part of the proposed IDS. However, it is understood from the literature that when clustering is preceded by the classification, it will be able to improve performance as it becomes

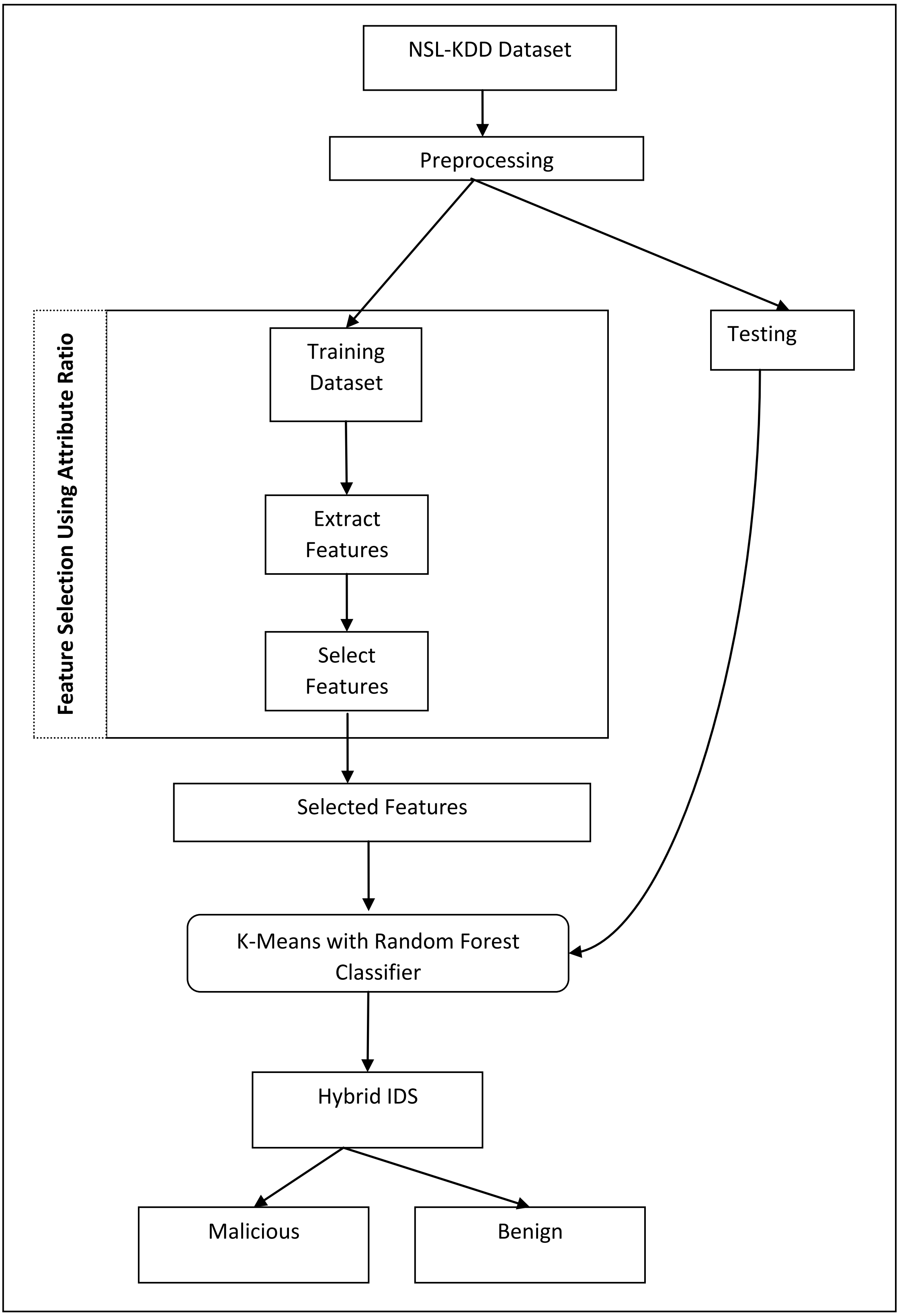

Figure 4: Hybrid IDS with RF and K-Means clustering

hybrid IDS. This is illustrated in Figure 4.

The approach of the hybrid IDS is similar to that of the one illustrated in Figure 3.

The difference is that the hybrid IDS is made up of both clustering and classification techniques. The clustering is made with K-Means while the classification is made with RF. However, the feature selection and other functions remain the same. In this project, many performance metrics are used for evaluating the IDS developed using machine learning algorithms. Confusion matrix is the basis for the metrics.

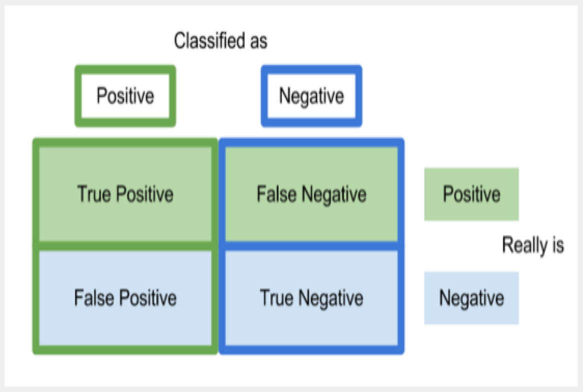

Figure 5: Confusion matrix

As presented in Figure 8, based on the correct and wrong predictions of intrusions, the initial measures like True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) are used for deriving other metrics like precision, recall, F1 Score and accuracy as provided in Eq. 1, Eq. 2, Eq. 3 and Eq. 4.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (3)$$

$$F1\ score = 2 * \frac{(Precision*Recall)}{(Precision+Recall)} \quad (4)$$

F1-Score is the measure which reflects the accuracy of the IDS. This measure is the harmonic mean of two measures like precision and recall.

4. EXPERIMENTAL RESULTS

The environment used for coding is Anaconda which is Python based data science platform with many IDEs available. The language used is Python, machine learning toolkit used in Scikit-learn and the distributed programming framework used is known as PySpark. Dataset used in this project is NSL-KDD which is acquired from [20].This section provides experimental results when IDS is used with K-Means and RF individually and also when they are combined with a hybrid IDS.

Probability Threshold	Precision	Recall	F1-Score	Accuracy
0.05	1	0.99	0.99	0.99
0.005	0.91	0.86	0.89	0.9

Table 1: Shows the performance of IDS with K-Means

As presented in Table 1, the performance of IDS with K-Means is provided against different probability thresholds like 0.05 and 0.005.

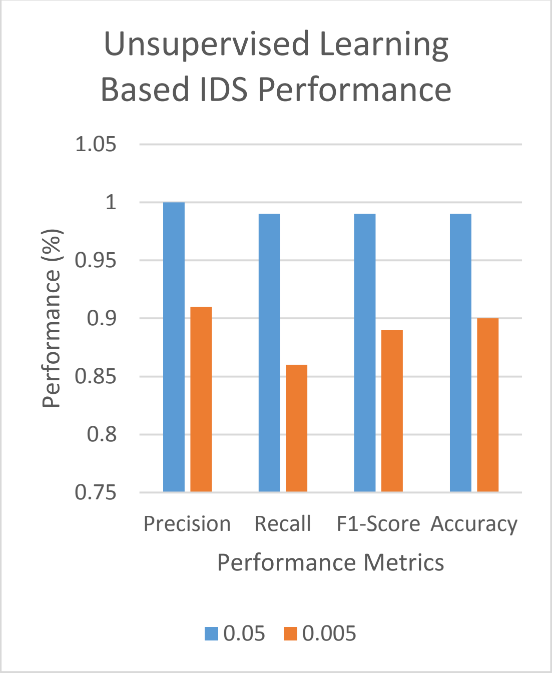

Figure 6: Performance of IDS with K-Means (Unsupervised Learning)

As presented in Figure 6, the performance metrics like accuracy, F1-score, recall and precision are provided in horizontal axis. The performance of the IDS when probability is set at 0.05 and 0.005 is presented in vertical axis. The performance of the IDS differs when probability threshold is changed. More probability threshold has higher performance when compared with the low probability threshold.

Probability Threshold	**Detection rate**	**False alarm rate**
0.05	0.997	0.149
0.005	0.91	0.86

Table 2: Shows the performance of IDS with K-Means

As presented in Table 2, the performance of IDS with K-Means is provided against different probability thresholds like 0.05 and 0.005.

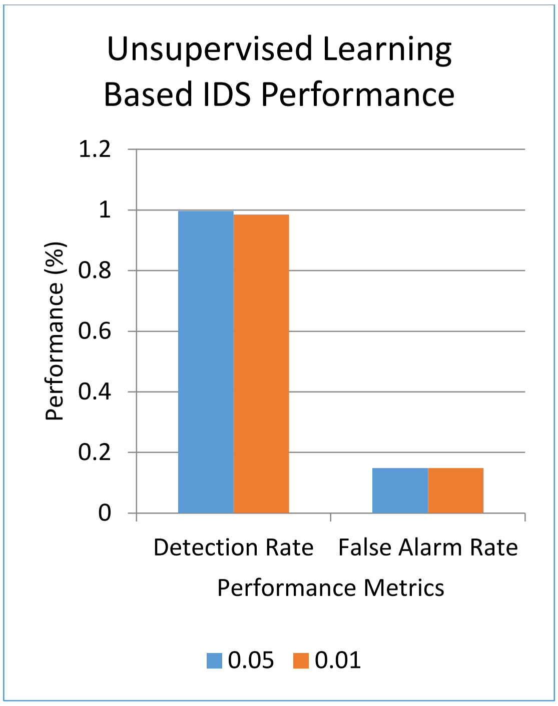

Figure 7: Performance of IDS with K-Means (Unsupervised Learning)

As presented in Figure 7, the performance metrics like detection rate and false positive rate are provided in horizontal axis. The performance of the IDS when probability is set at 0.05 and 0.01 is presented in vertical axis. The performance of the IDS differs when probability threshold is changed. However, less different is found in performance when there is less change in the probability threshold.

Probability Threshold	Precision	Recall	F1-Score	Accuracy
0.05	1	0.99	0.99	0.99
0.005	1	0.98	0.99	0.987

Table 3: Shows the performance of IDS with RF

As presented in Table 3, the performance of IDS with RF is provided against different probability thresholds like 0.05 and 0.005.

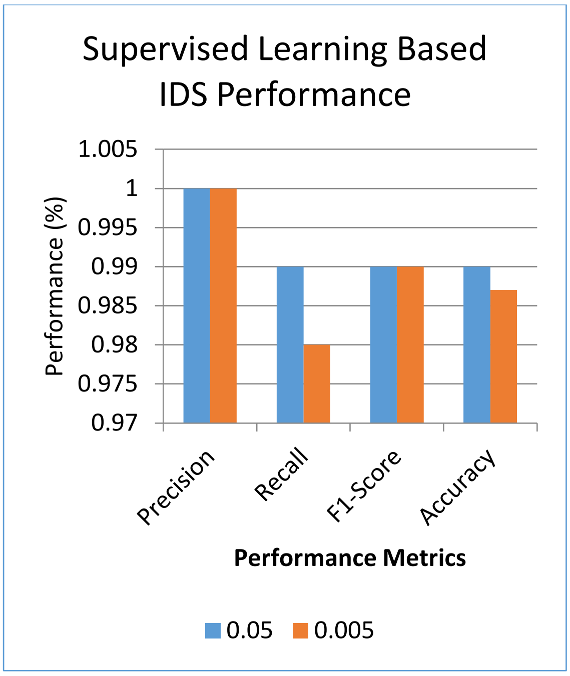

Figure 8: Performance of IDS with RF (Supervised Learning)

As presented in Figure 8, the performance metrics like accuracy, F1-score, recall and precision are provided in horizontal axis. The performance of the IDS when probability is set at 0.05 and 0.005 is presented in vertical axis. The performance of the IDS differs when probability threshold is changed. More probability threshold has higher performance when compared with the low probability threshold.

Probability Threshold	detection rate	False alarm rate
0.05	0.999	0.009
0.005	0.9353	0.1384

Table 4: Shows the performance of IDS with RF

As presented in Table 4, the performance of IDS with RF is provided against different probability thresholds like 0.05 and 0.005.

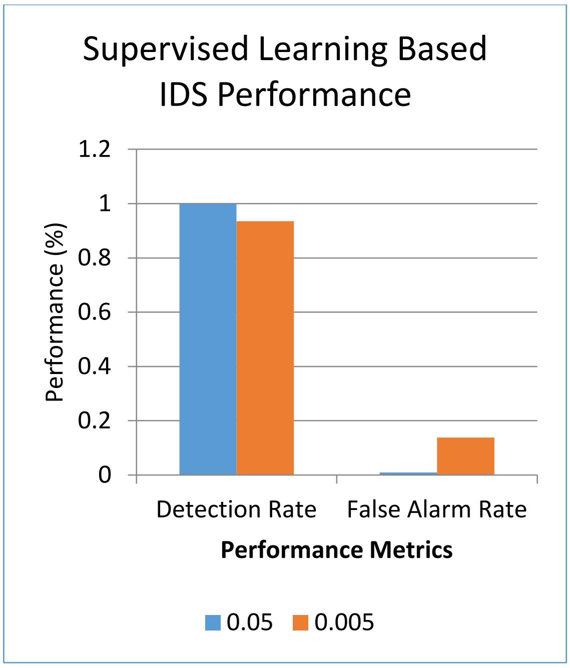

Figure 9: Performance of IDS with RF (Supervised Learning)

As presented in Figure 9, the performance metrics like detection rate and false positive rate are provided in horizontal axis. The performance of the IDS when probability is set at 0.05 and 0.005 is presented in vertical axis. The performance of the IDS differs when probability threshold is changed. However, less different is found in performance when there is less change in the probability threshold.

IDS	**Detection Rate**	**False Alarm Rate**	**F1 Score**
Hybrid IDS	0.99	0.14	0.94

Table 5: Shows the performance of IDS with K-Means and RF (Hybrid IDS)

As presented in Table 5, the performance of IDS with K-Means and RF is provided. The performance is shown in terms of detection rate, false alarm rate and F1-score.

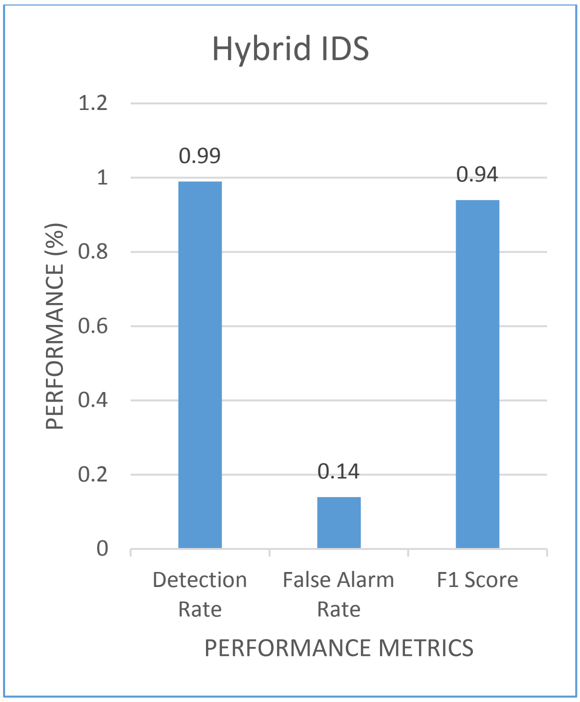

Figure 10: Performance of Hybrid IDS with K-Means and RF (Supervised Learning)

As presented in Figure 10, the performance metrics like detection rate, F1-score and false positive rate are provided in horizontal axis. The performance of the IDS is presented in vertical axis. The hybrid IDS showed 99% detection rate and 14% false alarm rate besides 94% F1-score.

5. CONCLUSION AND FUTURE WORK

Hybrid IDS is developed using machine learning approaches. It combines Random Forest classification and K-Means clustering. This will use both misuse detection and anomaly detection for improving performance of the IDS. These algorithms are evaluated for the four categories of attacks based on accuracy, false-alarm-rate, and detection-rate etc. The four kinds of intrusions considered are Denial of Service (DoS), trying to have unauthorized access such as guessing password (R2L), unauthorized access such as violating privileges (U2R) and probing (such as port scanning and surveillance). Experiments are made with NSL-KDD dataset. The IDS is supported by feature selection method known as attribute ratio. Anaconda data science platform issued to develop IDS. The results showed that the hybrid IDS has around 99% detection rate.

REFERENCES

[1] *Om, H., & Kundu, A. (2012). A hybrid system for reducing the false alarm rate of anomaly intrusion detection system. 2012 1st International Conference on Recent Advances in Information Technology (RAIT). P1-6.*

[2] Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2015). CANN: An intrusion detection system based on combining cluster centres and nearest neighbours. Knowledge-Based Systems, 78, 13–21.

[3] Maglaras, L. A., & Jiang, J. (2014). OCSVM model combined with K-means recursive clustering for intrusion detection in SCADA systems. 10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness. P1-2.

[4] Al-Yaseen, W. L., Othman, Z. A., & Nazri, M. Z. A. (2017). Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Systems with Applications, 67, 296–303.

[5] Ravale, U., Marathe, N., & Padiya, P. (2015). Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and RBF Kernel Function. Procedia Computer Science, 45, 428–435.

[6] Muniyandi, A. P., Rajeswari, R., & Rajaram, R. (2012). Network Anomaly Detection by Cascading K-Means Clustering and C4.5

Decision Tree algorithm. Procedia Engineering, 30, 174–182.

[7] 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.

[8] Gupta, G. P., & Kulariya, M. (2016). A Framework for Fast and Efficient Cyber Security Network Intrusion Detection Using Apache Spark. Procedia Computer Science, 93, 824–831.

[9] Chauhan, H., Kumar, V., Pundir, S., & Pilli, E. S. (2013). A Comparative Study of Classification Techniques for Intrusion Detection. 2013 International Symposium on Computational and Business Intelligence. P1-4.

[10] Thaseen, S., & Kumar, C. A. (2013). An analysis of supervised tree based classifiers for intrusion detection system. 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering. P1-6.

[11] Farnaaz, N., & Jabbar, M. A. (2016). Random Forest Modeling for Network Intrusion Detection System. Procedia Computer Science, 89, 213–217.

[12] Singh, K., Guntuku, S. C., Thakur, A., & Hota, C. (2014). Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests. Information Sciences, 278, 488–497.

[13] Mazini, M., Shirazi, B., & Mahdavi, I. (2018). Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms. Journal of King Saud University - Computer and Information Sciences. p1-30.

[14] Nadiammai, G. V., & Hemalatha, M. (2014). Effective approach toward Intrusion Detection System using data mining techniques. Egyptian Informatics Journal, 15(1), 37–50.

[15] Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, 25, 152–160.

[16] Aslahi-Shahri, B. M., Rahmani, R., Chizari, M., Maralani, A., Eslami, M., Golkar, M. J., & Ebrahimi, A. (2015). A hybrid method consisting of GA and SVM for intrusion detection system. Neural Computing and Applications, 27(6), 1669–1676.

[17] Maleh, Y., Ezzati, A., Qasmaoui, Y., & Mbida, M. (2015). A Global Hybrid Intrusion Detection System for Wireless Sensor Networks. Procedia Computer Science, 52, 1047–1052.

[18] Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 41(4), 1690–1700.

[19] Pan, S., Morris, T., & Adhikari, U. (2015). Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems. IEEE Transactions on Smart Grid, 6(6), 3104–3113.

[20] NSL-KDD Dataset. Retrieved from https://www.unb.ca/cic/datasets/index.html