**International Journal of Engineering Technology and Applied Science**

**(ISSN: 2395 3853), Vol. 6 Issue 10 October 2020**

# Pattern Matching based Anomaly Intrusion Detection using Machine Learning Supervised Algorithm

**[1]Pooja Tiwari, [2]R. N. Sharma**
**[1]M-Tech Scholar, [2]Assistant Professor**
**[12]Department of Computer Science Engineering, MPCT, Gwalior**

***Abstract-*** Intrusion detection systems play an important role in network security. Feature selection is the major challenging issue in IDS in order to reduce the useless and redundant features among the attributes. In this report, a hybrid learning approach through the combination of NAB (Numenta Anomaly Benchmark) and supervised classifier are proposed. In hybrid IDS we have used NAB algorithm for classification purpose. We took the help of matching and length searching training phase to reduce large heterogeneous dataset to a number of small homogeneous subsets. The proposed approach is compared and evaluated using KDD CUP 99 dataset. Only selected attributes are used for training formation as well as for classification purpose. As a result complexities get reduced and consequently performances get increased. Simulation results prove that accuracy rate and detection rate of DOS, PROBE, U2R and R2L increases by using our proposed method. Furthermore, for reduced feature attribute dataset performance of these hybrid IDS slightly increases as compare to all 41 attribute set. As well as the false alarm rate also decreases of proposed technique.
.

**KEYWORDS: ANN, KDD-99, Supervised Learning, AIS, NAB, Confidence Factor.**

## I INTRODUCTION

Now a day's development of any country or origination is depending upon its information technology system and all the information whether it's confidential, personal or public is shared through internet or network. So any country or organization needs to develop their information sharing network throughout the world with rapid speed. There is a rapid development in making such types of networks which available worldwide and have confidential information. But some time the intruder can attack over network where network based or client based firewall not capable enough to provide complete security against these types of threads.

Computer security is a very important issue to any or all users of pc systems. The rise of the web, pc attacks are increasing and may simply cause numerous dollar harm to a corporation. Detection of those attacks is a very important issue of pc security. Intrusion Detection Systems (IDS) technology is a good approach in addressing the issues of network security. The main objective of Intrusion Detection System is to observe unauthorized use, misuse and abuse of pc systems and laptop by each systems insiders and external intruders. There are many strategies following implement intrusion detection like statistical analysis knowledgeable systems, and state transition approaches etc., and these many approaches is based on the system were planned in recent years.

In order to provide complete security against these word wide thread IDS system play a key role. IDS system identifies the unauthorized activity that compromise the integrity, confidentially and availability of confidential information.

Conventional IDS is based on continuous monitoring of well know attack by their extensive knowledge of signature to detect intrusion. This method based on pattern recognitions of various audit streams and detect intrusion by comparing their pattern provide by human expert. The pattern has been manually revised for a new type of intrusion whenever discover. The basic limitation of this pattern based Method is cannot detect emerging cyber thread.

Artificial Immune System is an emerging technology in order to fine the intruders or making the IDS. Recently AIS is a new bio-inspired model, which is applied for solving various security problems in the field of information security, genetic algorithms, neural networks, evolutionary algorithms and swarm intelligence. As one of the solutions to intrusion detection problems, AIS have shown their advantages. To improve the correlation factor and minimizing the false alarm generation we used the concept of AIS and NAB Algorithm to identify the intrusion in the system.

Detection and prevention of anomaly over the internet in real time scenario is a big challenge. The versatile feature and dynamic nature of anomaly attack emerges the issue of discovery and avoidance of attack. The abnormality assault is umbrella of different assault such DOS, Probe, U2R, R2L and numerous mixes of assaults. For the discovery of abnormality assault utilized firewall, interruption identification framework, antivirus and numerous more application programming are used [1, 2, and 3]. The handling of identification is ease back because of vast number of interruption characteristic, now different creators utilized machine learning method and highlight diminishment prepare for the characterization and detection in intrusion detection system. Big data is the accumulation of extensive informational collections and it winds up plainly hard handling utilizing reasonable customary information preparing applications or database administration instruments. The difficulties incorporate

procuring, putting away, seeking, sharing, exchanging, breaking down and imagining. The pattern of big data is due to the other useful information that can be derived from analysis of large set of related data, allowing correlations to be made to spot business trends, prevent diseases, determine quality of research, link legal citations, contest crime, and find current roadway traffic conditions. In the circumstance of the data explosion phenomenon, existing performance models for Map-Reduce are applicable for specific production workloads, but are to reveal the real capabilities of the processing system under heavy workloads that process tens of terabytes of data. While processing a query in big data, speed is a significant demand. The combination of suitable index for big data and current preprocessing technology will be a desirable solution when we encounter this kind of problems. Big Data is used in many real-world areas such as telecommunications, health care, pharmaceutical or financial businesses. Machine learning offers various algorithm for classification, clustering and combination of clustering and classification. The clustering techniques provides various algorithm such as k-means, k-mod, FCM and many more algorithm. Instead of clustering the classification algorithm gives more accuracy in terms of detection.

## II SYSTEM METHODS

In this section we present the conclusion results of existing intrusion detection techniques for detection DOS attacks. Intrusion detection system in a very popular and computationally expensive task.

In this paper [1] Network intrusion detection is an important component for network management and it is defense mechanism for network security. A real-time network intrusion detection system has been presented in this work. This proposed Support Vector Machine based network intrusion detection system is evaluated with KDD 99 dataset. The proposed system is developed with the consideration of big streaming data. The experimental results show that the proposed system is feasible for stream processing of network traffic data for detection of network intrusion with high accuracy.

In this paper [2] Intrusion detection systems play an important role in network security. Feature selection is the major challenging issue in IDS in order to reduce the useless and redundant features among the attributes. In this report, a hybrid learning approach through combination of K - Means clustering and SVM classifier are proposed. In hybrid IDS we have used RBF kernel function of SVM for classification purpose. They took the help of K- Means clustering technique to reduce large heterogeneous dataset to a number of small homogeneous subsets. The proposed approach is compared and evaluated using KDD CUP 99 dataset. Only selected attributes are used for cluster formation as well as for classification purpose. As a result complexities get reduced and consequently performances get increased. Simulation results prove that accuracy rate and detection rate of DOS, PROBE, U2R and R2L increases by using this proposed method. Furthermore, for reduced feature attribute dataset performance of these hybrid IDS slightly increases as compare to all 41 attribute set. As well as the false alarm rate also decreases of proposed technique.

In this paper [3], they have presented an immune system inspired unsupervised intrusion detection system. Unlike other methods it assumes the coexistence of viruses and self, and therefore allows adaptability without supervision. It does not require labeled data, but learns the signatures of normal connections and intrusions from the dataset itself. It consists of two units: the T cell and the B-cell. The T-cells are formulated using a hidden Markov model, while additionally incorporating information from past observations to improve the model adaptively. The B-cells combine inputs from T-cells with broader-level feature information to weed out false alarms. Its two – layer structure, similar to biological immune systems, achieves a high detection rate and helps keep false alarm rate in check; while other models show either very low detection rates or unacceptably high false alarm rates. The algorithm can operate in real time. Since viruses mutate, the ability of an IDS to detect viruses previously not encountered is of paramount importance. Our model is able to do that, and represents a significant improvement over other models in this respect. This is evident from results for the ips-weep dataset, where our model successfully detected a type of attack it had not been exposed to in the training stage.

## IV PROPOSED METHODOLOGY

The planned design contains varied modules every outlined with a particular purpose and connected along to spot the precise unauthorized person within the given system.

The objective of the trainer module is to train the NN for classifying the log of real time network data with desired attributes as attack or normal. Some of the advantages of NN are:

1. They have the inherent property of learning through training.
2. The complex internal structures enable them to learn and accommodate large number of patterns.
3. They can generalize the knowledge acquired through training for similar patterns.
4. They have efficient storage capability for a large set of patterns.

The flow chart for the training & testing phase of the NAB module is as shown in Figure 1.

The sequence of steps is as follows.

1. Start the software module environment.
2. Select the data set which has to be trained by the training module from the main data base folder.
3. Fetch each and every attribute class and count. These three point include the training phase of our projected module.

4. Then construct the new data for new collected information.
5. Last but not the least of our project method module is update these information into the data set and gets trained data set.

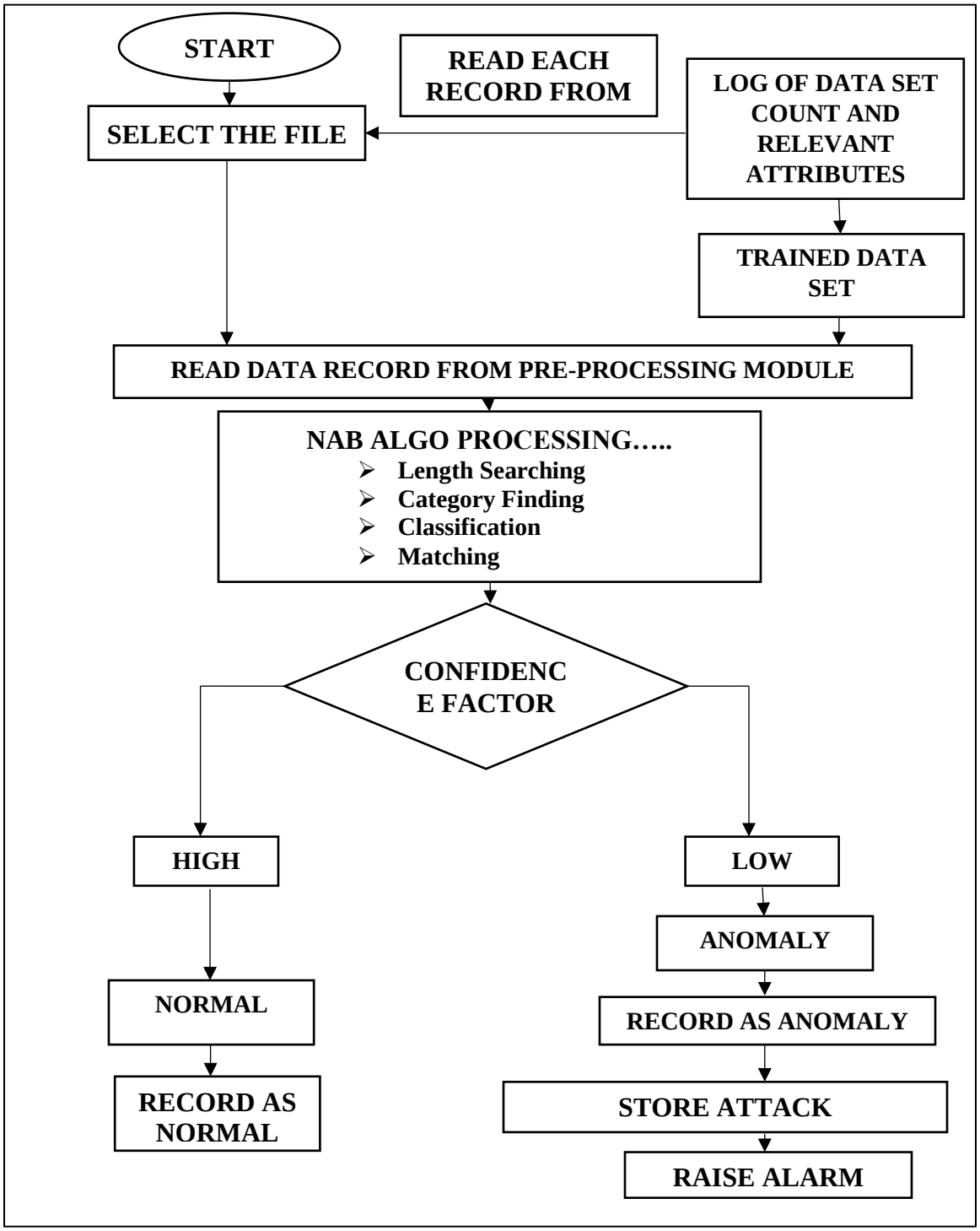


**Figure 1: Flow Chart for Detection Phase**

These characteristics of NN, motivated to choose it for classification as in the training phase both the input and output data set is available and supervised learning method can be used. In this phase, the system gathers knowledge about the normal behavior of the network users from the preprocessed input data, and stores the acquired

knowledge. In the detection phase, the system detects attacks based on the knowledge which is achieved during the training phase, and notify the system administrator.

**Step of figure 1 flow chart for detection phase:**

1. Start the software module environment.
2. Select the data set which has to be tested by the system module from the main data base folder. (Log of data set count and relevant attribute as referred a storage of data, in our case we have used KDD-99 as data set for both training and testing )
3. Read the each data record of tested data set and trained data set according to pre-processing module.
4. In this step first length searching of the data set which is form of string for example string length how many column and row contain the string of data.
5. Then next process is category finding in which we have find out the data set attributes belong to which category.
6. Then done classification according to their class like main class (normal, probe, DOS, R2L, L2R).
7. Then done matching process, in which we have made match between tested data set and training data set.
8. Next process find out confidence factor of the tested data set, if the probability uncertainty of data high then confidence factor low and if the probability of uncertainty of data low then confidence factor high.
9. High confidence factor treat as normal, and store record as normal. Low confidence factor treat as anomaly detection it means intrusion is present and it has recorded as anomaly and raise alarm.

**Algorithm for Detection Phase:** The algorithm which is using in our proposed methodology for detect the anomaly behavior of test data set discussed below. (Numenta Anomaly Benchmark [NAB]) algorithm used and elaborate in the following points.

***Step 1: Start***

- ***Select the test excel data sheet (KDD-Cup-99 Dataset)***
- ***Define the variable loop value to read the data sheet length***
- ***Format test data into string***
- ***Create new two arrays (final_res and final_act) of Zeros***
- ***Read the test excel data sheet***

***Step 2:***

- ***Load trained data set for matching***
- ***Define main category***
- ***Taking particular case i.e. case-14, case-23, case-otherwise***

***Pri_col = 14; (For case = 14)***
***Pri_col = 23; (For case = 23)***
***Pri_col = 44: (For case = otherwise)***

***Step 3: Define some variables,***
***Wrong = [ ]; cmp = [ ]; actual = [ ]; search = [1: Upto]; wn = [ ]; wa = [ ]***

***Step 4: Length Searching***
**for l = 1:length(search)**
**r_no = search(l)**
**read_data = data1(r_no,1:42);**

***Step 5: For desire category finding,***
***Comparing the two string [Test data (d) and trained match data]***

***Step 6: Finding Confidence Factor***
***If decision ==High, then (it is case of anomaly detection)***
***If decision ==Low, then (it is case of normal detection)***

***Step 7: Final Matching,***
***If match (length) then final_act == final_res;***
***If final_act ≠ final_res, then wrong predicted;***

***Step 8: For Normal, the value of final_res = final_act = 1***
***If value of final_act == final_res =2, (Detected as DOS)***
***If value of final_act == final_res =3, (Detected as R2L_U2L)***
***If value of final_act == final_res =4, (Detected as Probe)***

## V SIMULATION RESULTS

The data set employed in the experiments is ‘‘KDD Cup 1999 Data’’, that may be a subversion of Defense Advanced (Defense Advanced Research Projects Agency) 1998 dataset.

The KDD cup ninety nine dataset Includes a group of forty one options derived from every association and a label that specifies the standing of association records as either traditional or specific attack sort. These options had all types of continuous and symbolic with extensively variable ranges falling in four categories:

Proposed Work has been implemented in MATLAB 2014b framework. Figure 2, shows the main window of proposed IDS system. In this main window of proposed IDS system.

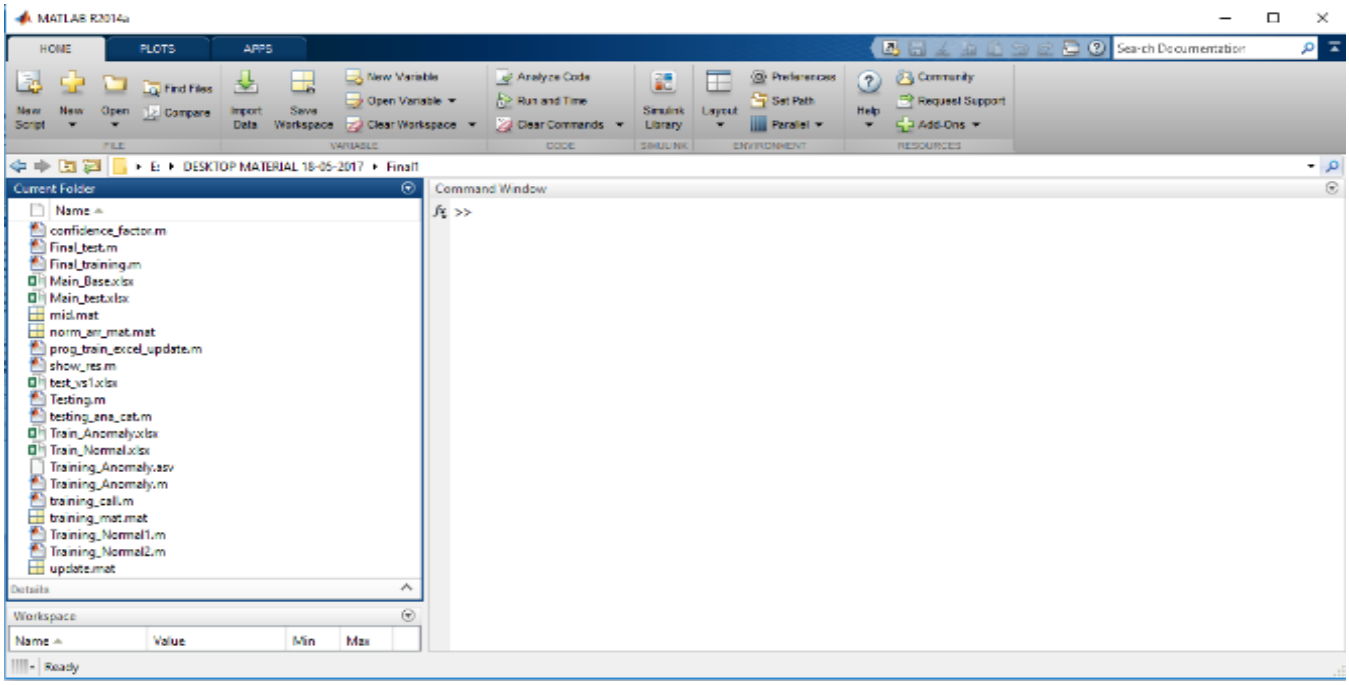


**Figure 2 Show the Main MATLAB Window Environment**

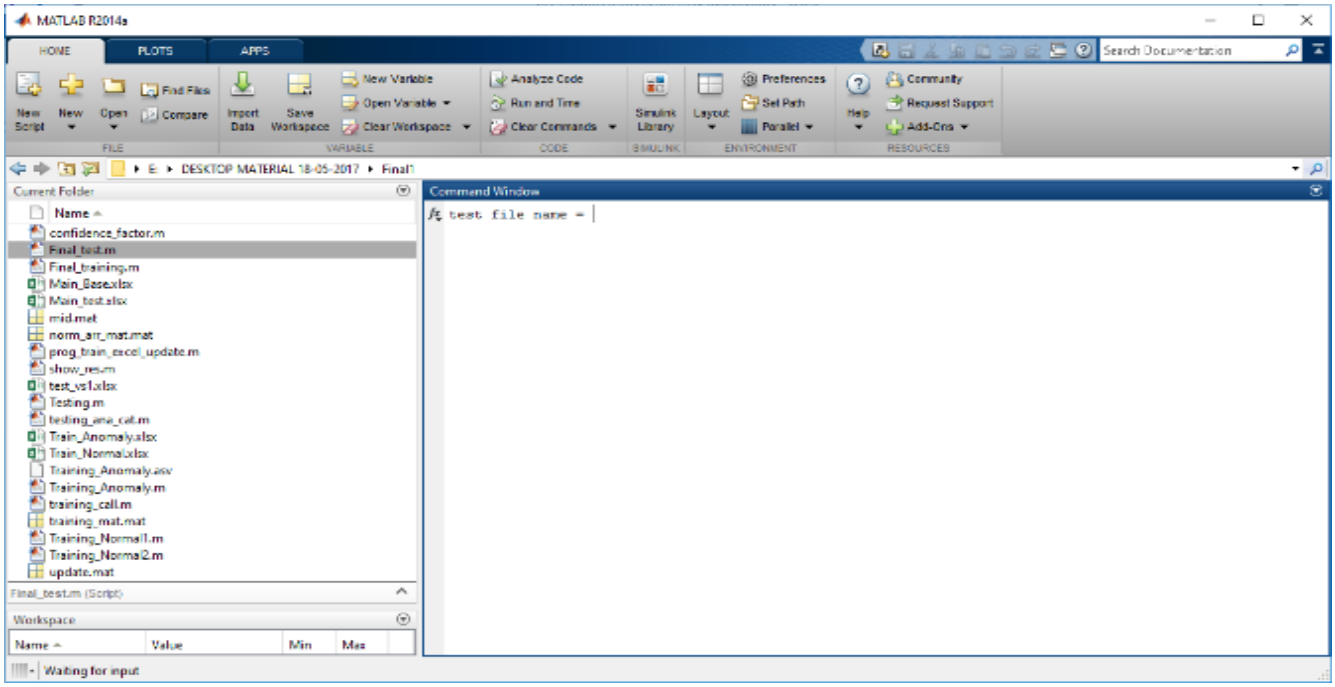


**Figure 3 Shows the Training Phase of Proposed Algorithm in MATLAB Environment**

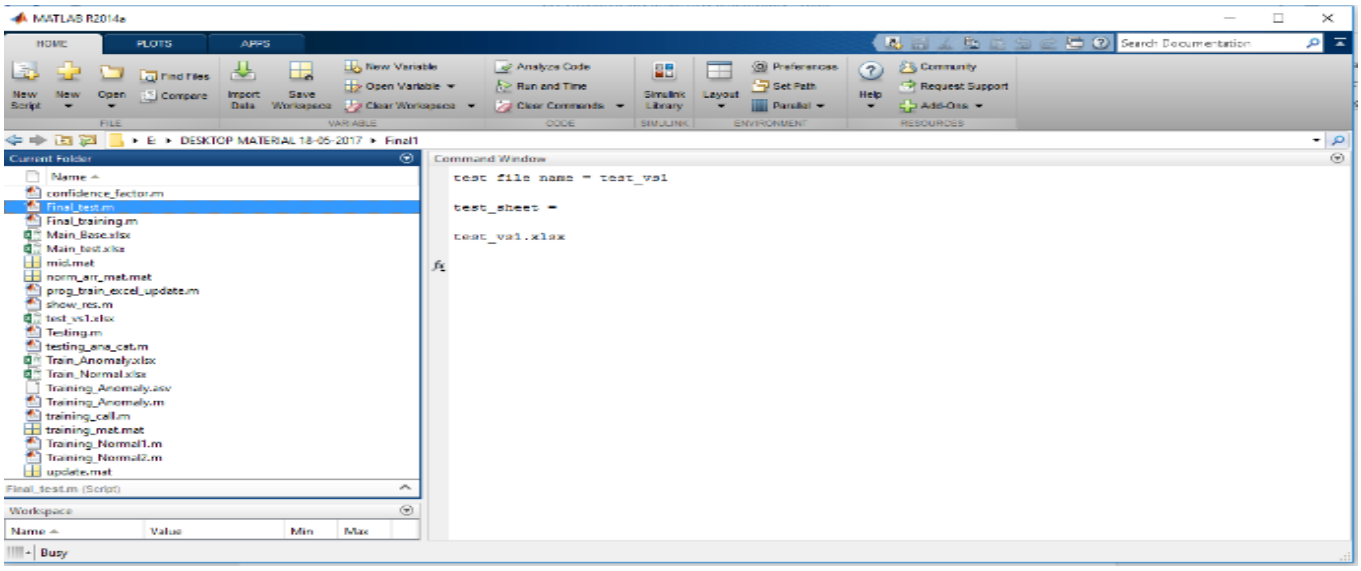


**Figure 4 Shows the Testing Phase of Proposed Methodology for Test File**

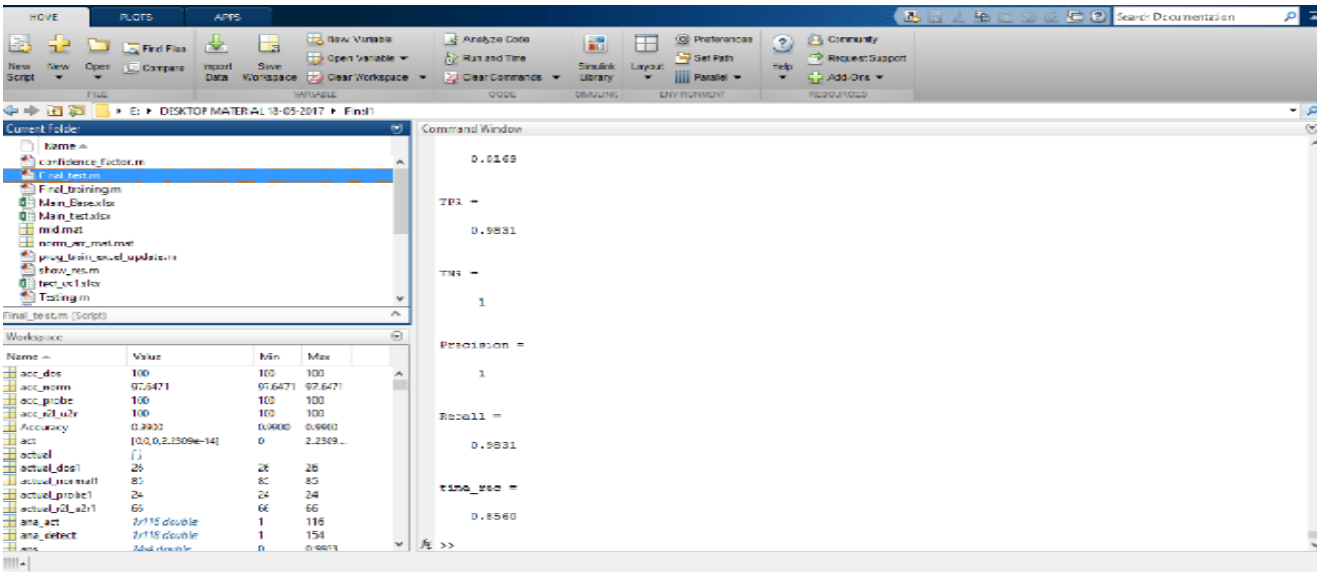


**Figure 5 Shows the Results of Our Proposed Methodology**

Figure 4 shows that the testing phase of test data file of proposed methodology. In this process basically find the abnormalities about the test data set. These finding basically distinguished between normal and abnormal behavior of our data set. The sequences that our proposed algorithm follows: The searching according to length and location, Matching process, and determine the confidence factor and then find out the detection rate and accuracy of the system.

Figure 3 show the training phase of proposed algorithm in MATLAB environment. The training phase of the test file follows some tasks.

Figure 5 show that the results of our proposed methodology. In this figure shows the accuracy results for all attributes class data set.

## VI COMPARISON WITH EXISTING MECHANISM (ANOMALY)

This section shows that the comparison in term of accuracy and detection rate with the existing mechanism. In this section as existing mechanism we shows the results of the system using numerous algorithm like K-means clustering, support vector machine and k-means with RBF kernels of SVM**.**

**1. Accuracy Results (In Percentage) for All Attribute Set**

**Table 1 Accuracy Results (In Percentage) for All Attribute Set**

| METHOD / DATA-SET | KMSVM | KM | SVM | (PROPOSED) |
|---|---|---|---|---|
| DOS | 93.33 | 86.67 | 40 | 98.50 |
| PROBE | 100 | 87.50 | 75 | 100 |
| U2L | 93.75 | 68.75 | 62.50 | 93.78 |
| R2L | 87.50 | 75 | 68.75 | 90.13 |
| ALL | 80.28 | 73.24 | 49.30 | 90.07 |

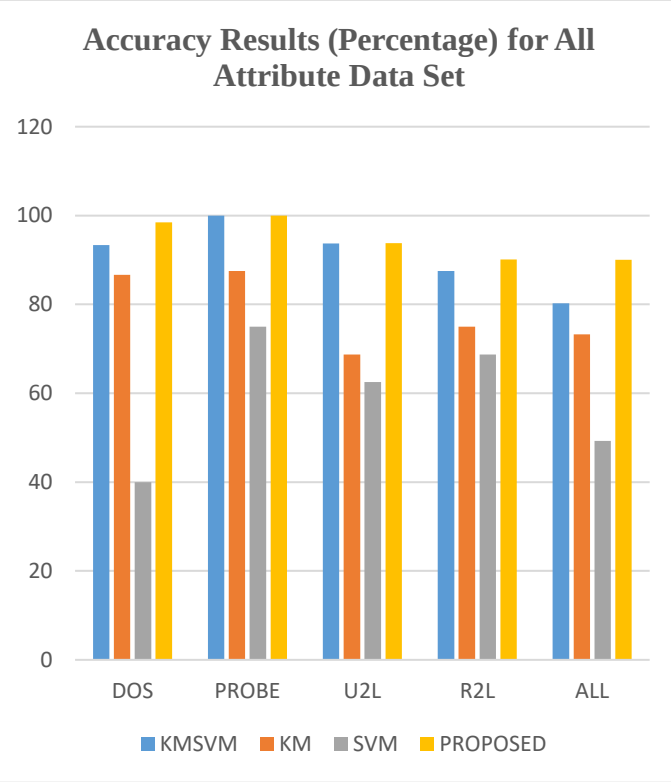


**Figure 6 Chart Shows the Accuracy Results (In Percentage) for All Attribute Data Set**

In table 1 shows the accuracy (percentage) value of different algorithm used for IDS, in this table we describe the algorithm used system accuracy k-means support vector machine (KMSVM), K-Means Clustering (KM) and Support Vector Machine (SVM) and proposed hybrid model. Also we plot these value of accuracy of different algorithm used IDS system for all attribute class data set in figure 6.

**2. Accuracy Results (In Percentage) for Reduced Attribute Set**

**Table 2 Accuracy Results (In Percentage) for Reduced Attribute Set**

| METHOD / DATA-SET | KMSVM | KM | SVM | (PROPOSED) |
|---|---|---|---|---|
| DOS | 100 | 100 | 57.98 | 100 |
| PROBE | 91.30 | 91.30 | 73.91 | 100 |
| U2L | 100 | 77.78 | 88.89 | 100 |
| R2L | 94.12 | 82.35 | 70.59 | 98.50 |
| ALL | 87.01 | 76.62 | 62.34 | 93.34 |

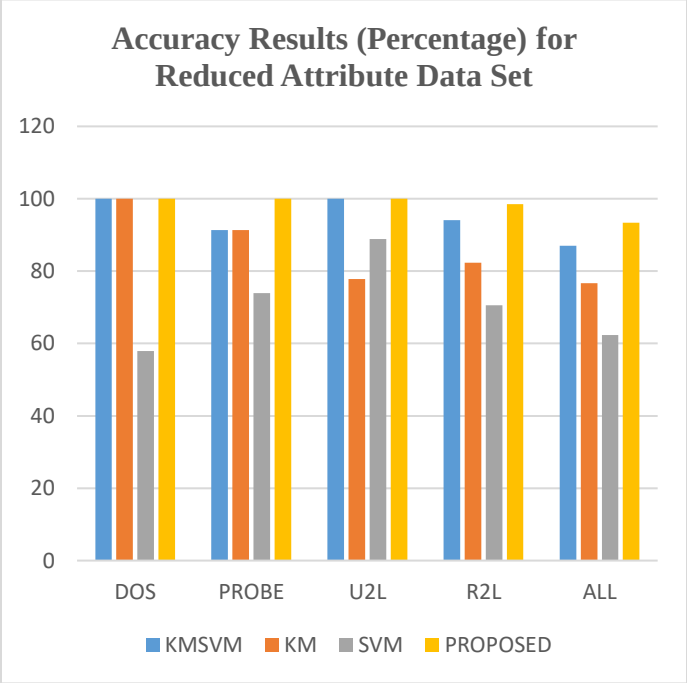


**Figure 7 Chart Shows the Accuracy Results (In Percentage) for Reduced Attribute Data Set**

In table 2 shows the accuracy (percentage) value of different algorithm used for IDS, in this table we describe the algorithm used system accuracy k-means support vector machine (KMSVM), K-Means Clustering (KM) and Support Vector Machine (SVM) and proposed hybrid model. Also we plot these value of accuracy of different algorithm used IDS system for all attribute class data set in

figure 7. Also Figure 6 and Figure 7 shows comparison of the simulation result. It gives the comparison of the degree of Accuracy rate of IDS system by using traditional classification method namely KMSVM, SVM and KM with our proposed method Hybrid model. Hybrid modal increases the accuracy rate by encapsulating B and T-Cell theory with NAB algorithm. As shows in figure 6 KMSVM, KM and SVM classification algorithm alone having accuracy rate for all attributes data set never reaches even 93.75% whereas hybrid model having accuracy rate up to 94.50%. As shows in figure 7, KMSVM, KM and SVM classification algorithm alone having accuracy rate for reduced attributes data set reaches even average accuracy 97% whereas hybrid model having accuracy rate up to 98.50%.

### 3. Detection Rate Results (In Percentage) for All Attribute Set

**Table 3 Detection Rate Results (In Percentage) for All Attribute Set**

| METHOD / DATA-SET | KMSVM | KM | SVM | (PROPOSED) |
|---|---|---|---|---|
| DOS | 91.67 | 85.71 | 100 | 99.30 |
| PROBE | 100 | 89.47 | 75 | 98.70 |
| U2L | 90.91 | 66.67 | 62 | 95.80 |
| R2L | 84.62 | 73.33 | 100 | 100 |
| ALL | 79.37 | 72.86 | 100 | 99.80 |

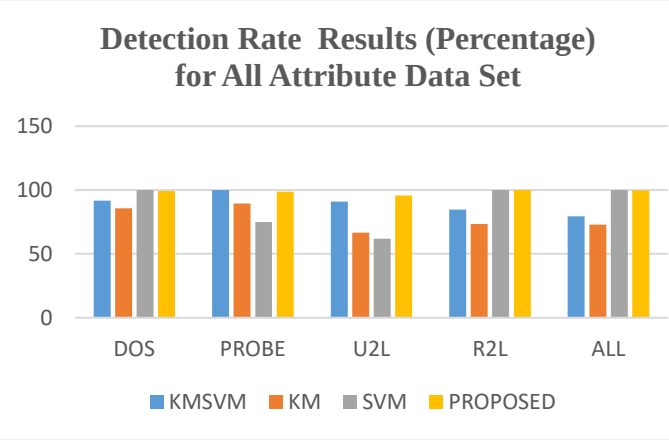


**Figure 8 Chart Shows the Detection Results (In Percentage) for All Attribute Data**

In table 3 shows the detection rate (percentage) value of different algorithm used for IDS, in this table we describe the algorithm used system accuracy k-means support vector machine (KMSVM), K-Means Clustering (KM) and Support Vector Machine (SVM) and proposed hybrid model. Also we plot these value of detection rate of different algorithm used IDS system for all attribute class data set in figure 8.

## VII CONCLUSION

In order to overcome all these deficiency from IDS, system over network, we propose a novel dual detection of IDS based on AIS that integrating the NAB algorithm. The training phase helps us to solve the problem of correlation and NAB theory resolves the problem of unknown and rapidly evolving harmful attacks. The simulation results shows that the proposed method has improved the accuracy rates, minimizing false +ve and false –ve alarm generation and to increase the efficiency and accuracy of the IDS system.

## REFERENCES

1) Baojiang Cui and Shanshan He "Anomaly detection model based on Hadoop platform and Weka interface", Innovative Mobile and Internet Services in Ubiquitous Computing, 2016, Pp 84-89.
2) Sherenaz Al-Haj Baddar, Alessio Merlo and Mauro Migliardi "Anomaly Detection in Computer Networks: A State-of-the-Art Review", Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 2015, Pp 29-64.
3) Zhengbing Hu, SergiyGnatyuk, Oksana Koval, Viktor Gnatyuk and SerhiiBondarovets "Anomaly Detection System in Secure Cloud Computing Environment", I. J. Computer Network and Information Security, 2017, Pp 10-21.
4) Muhammad Asif Manzoor, Yasser Morgan, "Real-time Support Vector Machine Based Network Intrusion Detection System Using Apache Storm", IEEE 2016.
5) Prof. Ujwala Ravale, Prof. Nilesh Marathe, Prof. Puja Padiya, "Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and RBF Kernel Function", Elsevier 2015.
6) Manjari Jha, Raj Acharya, "An Immune inspired Unsupervised Intrusion Detection System for Detection of Novel Attacks", IEEE 2016.