# An Implementation of Model using Machine Learning Algorithm for Intrusion Detection System

**Namrata Pandey[1], Dr. Pawan Kumar Patnaik[2], Mr. Sargam Gupta[3]**

[1-3]*Department of CSE, B.I.T. Durg Bhilai, Chhattisgarh, India*

---

***Abstract*— With the limitation of traditional technologies like firewalls and due to the advancement in the era of technologies the network security are on high risk, which further emerges the need of new technologies and more advanced solutions for cyber security. Previous studies have showcase many Intrusion detection systems which are not very capable of identifying and classifying the attacks present in the network like DoS(Denial of Service), Probe, U2R(User to Root) and R2L(Remote to Local) . For better evaluation and identification, this paper proposed a model based on machine learning algorithm for Intrusion detection system. Model is built using the standard data sets which is highly preferred for intrusion detection system is *Knowledge Discovery in Databases* or KDD for short.

***Keywords*— Network Intrusion, Machine Learning, attacks, features, KDD datasets, Network.**

## 1. INTRODUCTION:

Artificial intelligence (AI) is an evergreen branch in field of computer science. A subset branch of Artificial Intelligence namely Machine learning playing a vital role in today's trend of technology and development by contributing in the field of automation. Machine learning provides system (model) the ability to learn by its own without human intervention. Machine learning is not limited to particular domain in fact has a significant contribution in the field with a wide range of application area like health care, educational sector, research centres, cyber security and many more.

As aforementioned that one of the application areas of machine learning is cyber security, we can apply different machine learning algorithms to design model that can be used as intrusion detection system which is one of the biggest treat in cyber security. In this paper we are constructing a model using machine learning supervised algorithm and the KDD data sets which is one of the highly preferred datasets by the researchers in field of cyber security[5][6].

In this research paper we are making use of only few features and applying the random forest learning algorithm. As the number of feature used in the model affects the performance of model as well as speed of computation and resource [1].

The redundancy of features has been removed using the feature scaling techniques.

The entire rest of the paper illustrates the sections as follows. Section II presents the research methodology comprises of various machine learning steps. Section III presents experimental results and discussion of the model. And the last section IV gives the conclusion work carried out in the research work.

## 2. RESEARCH METHODOLOGY

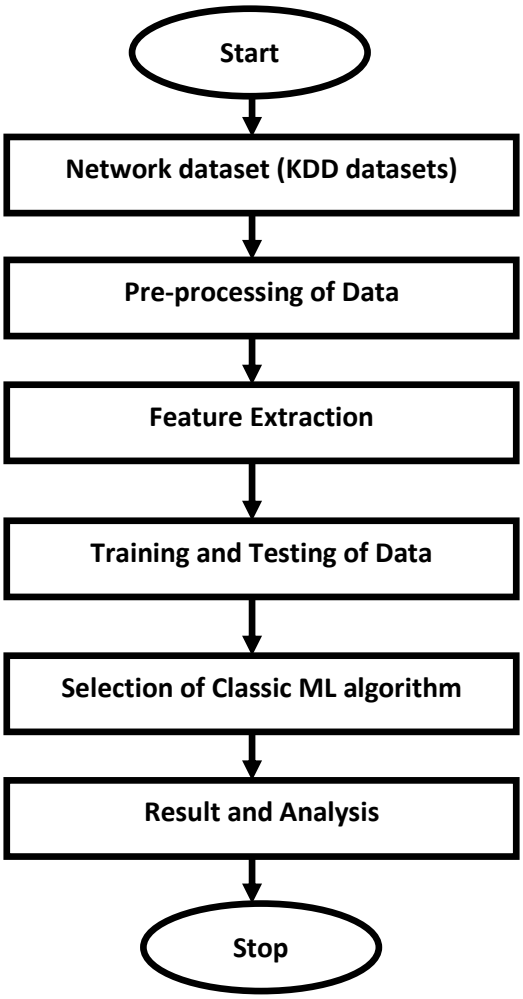


Fig (1) Methodology Steps

### 2.1 *Network dataset (KDD datasets)*

Knowledge Discovery in Databases or KDD for short is datasets used for intrusion detection system [2]. KDD data sets are highly processed datasets and are available in different version with wide varieties of attributes used in detecting the attacks which are tabulated in tabular format, where sets of features representing the dataset and the values assigned to the features represents the instance[4][5].

### *2.2 Pre-processing of Data*

One of the most important and core process carried out during designing any machine learning model is the pre-processing step [1][2]. Pre-processing is the heart of any machine learning model implementation of this process helps in designing the anomaly free model. This process is carried out by performing few more steps like sampling the datasets, cleaning the datasets and dealing with missing values.

#### *2.2.1 Sampling*

Processing the entire datasets is quite expensive and time consuming to reduce this complexity and expense data sampling is carried out. Sampling is applied over the entire population of datasets, there after selecting few sets of data ensuring that the selected datasets will obtain the result that is close to the one applied to the entire datasets. The commonly used sampling techniques have been illustrated below:

*a. Random sampling:* Random sampling is the highly preferred sampling technique by researchers. In random sampling Uniform selection of datasets take place. In other words, we can say that in a set of data size n, the probability of being selected for all instances is equal that is 1.

*b. Sampling* based on with replacement and without replacements: In replacement sampling selection of instance can be more than one time whereas in case of sampling without replacement, instances are removed from the selection pool once selected.

*c. Stratified sampling:* Other type of sampling commonly used is stratified sampling technique, in which data sets are segmented into n bins; after the segmentation using random sampling technique fixed number of instances are selected.

### *2.2.2 Removal of Noise*

One of the most important steps of data pre-processing is the removal of noise from the data sets. As presence of noise causes distortion and may lead to adverse affect in the performance of the models. For the removal of noise different filtering algorithms are applied.

### *2.2.3 Missing Values*

In a record of features many values are missing which is considered as missing values, which further leads to many complexities while designing a model using these values. So in order to reduce such complexity these values must be treated using any method as presented below:

1. Treating the missing values by removing the instances holding missing values.
2. Secondly we can place some estimated values in place of missing values.
3. Thirdly running the data mining algorithms by ignoring the missing values.

1.1 Feature Extraction

Feature Extraction is again the most important step which is used to eliminate the redundant and irrelevant data from the datasets [5]. Feature selection helps in selecting the limited features which are helpful in obtaining the model desired from all the features. Large sets of feature leads to time complexity as well as computation complexity thus to reduce this complexity feature Extraction is implemented.

### *2.3 Training and testing of Data*

The entire dataset which we used to train our model is divided into two sets one set is used as training set which is used for training our model and the other set is used as testing set which is used to test the model[3][4]. While dividing the entire datasets into two subsets following key points should be concerned:

1. Division of the datasets should be done in such a way that training sets should hold more data compared to testing sets.
2. Avoid using the same datasets for both testing and training.

*2.5 Selection of Classic Machine learning algorithm*

The selection of machine learning algorithms highly depends upon the nature of the problem statements.

## 3. RESULT AND DISCUSSION

The accuracy of the model is improved while reducing the number of features and the performance has also been improved. Model is trained using the random forest learning algorithm which gives better accuracy compared to the other learning model like logistic, svm. In this paper we used pre-processing step for removal of null or missing value and sampling is done. The accuracy Test for different types of attacks has been illustrated in table.1 below.

**Table I. Accuracy Test for different Types of attacks**

| Classification Algorithm | Class Name | Test Accuracy (%) with 40 Features Test | Accuracy (%) with 14 Features |
|---|---|---|---|
| Random Forest | **Normal** | **98.2** | **99.2** |
| | **DOS** | **97.1** | **98.0** |
| | **Probe** | **96.4** | **98.3** |
| | **R2L** | **96.6** | **97.9** |
| | **U2R** | **95.3** | **97.5** |

## 4. CONCLUSION

This study demonstrates an efficient approach to detect and identify network intrusion using machine learning algorithms. Standard data set KDD data set has been used for building the model. Many other algorithms has been used which are effective in the identification and classification of the attack types like denial of service, user to remote etc. Features and values of features play an important role in designing the model as it directly affects the performance and complexity of the model. As we can see in the table.1 that reduction in the number of the features increases the accuracy of the model.

## 5. REFERENCES

[1] Wattanapongsakorn, N., Sangkatsanee, P., Srakaew, S., & Charnsripinyo, C. (2011, September). Classifying network attack types with machine learning approach. In 7th International Conference on Networked Computing (pp. 98-102). IEEE

[2] Kumar, S., Viinikainen, A., & Hamalainen, T. (2016, December). Machine learning classification model for network based intrusion detection system. In 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST) (pp. 242-249). IEEE.

[3] Revathi, S., & Malathi, A. (2013). A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection. International Journal of Engineering Research & Technology (IJERT), 2(12), 1848-1853.

[4] Park, K., Song, Y., & Cheong, Y. G. (2018, March). Classification of attack types for intrusion detection systems using a machine learning algorithm. In 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 282-286). IEEE.

[5] M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), pp. 172–179, 2003

[6] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup 99/kddcup99.html, October 2007.