# INTRUSION DETECTION USING FEATURE SELECTION AND MACHINE LEARNING ALGORITHM WITH MISUSE DETECTION

Harvinder Pal Singh Sasan and Meenakshi Sharma

Department of Computer Engineering, Sri Sai College of Engineering and Technology, Punjab, India

***ABSTRACT***

*In order to avoid illegitimate use of any intruder, intrusion detection over the network is one of the critical issues. An intruder may enter any network or system or server by intruding malicious packets into the system in order to steal, sniff, manipulate or corrupt any useful and secret information, this process is referred to as intrusion whereas when packets are transmitted by intruder over the network for any purpose of intrusion is referred to as attack. With the expanding networking technology, millions of servers communicate with each other and this expansion is always in progress every day. Due to this fact, more and more intruders get attention; and so to overcome this need of smart intrusion detection model is a primary requirement.*

*By analyzing the feature selection methods the identification of essential features of NSL-KDD data set is done, then by using selected features and machine learning approach and analyzing the basic features of networks over the data set a hybrid algorithm is made. Finally a model is produced over the algorithm containing the rules for the network features.*

*A hybrid misuse intrusion detection model is made to find attacks on system to improve the intrusion detection. Based on prior features, intrusions on the system can be detected without any previous learning. This model contains the advantage of feature selection and machine learning techniques with misuse detection.*

***KEYWORDS***

*Feature Selection, NSL-KDD data set, Accuracy, Misuse Detection, Intrusion Detection.*

## 1. INTRODUCTION

Attacks in the system are becoming frequent and their detection is gaining importance. With the advancement of technology and dependence of human on growing internet and its applications, the safety of information, data flowing over the networks is becoming crucial. In order to prevent this crucial information and to achieve confidentiality, security of networks is one of the critical requirement of the growing network world these days [1].

In the current scenario intrusion detection is mostly human dependent, human analysis is required for detection of intrusion [2]. Systems depend heavily on manual input.

Intrusion detection is today important for business organization, system applications and for large number of servers and on-line services running in the system. But, for preventing the data over the network, enhancing the efficiency of the intrusion detection model is also equally important [3].

There are mainly two types of Intrusion detection system, named as Network based and Host based.

### Network Based Intrusion Detection System (NIDS)

The prominence factor of network based intrusion detection system (NIDS) is that at any single instance of time, NIDS can monitor multiple systems in a network in parallel. NIDS finds its best values when each single packet is analyzed that is about to move into the network through the firewall implemented above the network and thus helps in monitoring the information traversing through the network and detects any adversary or intrusion activities.

### Host Based Intrusion Detection System (HIDS)

In contrast to NIDS, host based intrusion detection system (HIDS) monitors the activities of an individual host or computer system. The primary focus of host based intrusion detection system is on the operating system activities and events. However, in network systems also, HIDS finds its best values in finding the flow of information and detecting the attacks over the network based on the events occurred within the network.

### 1.1. Analysis Approaches

The two main categories through which network can be analyzed for the detection of intrusion are Misuse detection and Anomaly detection.

### Misuse Detection

Misuse detection is an approach where the detection of intrusions is based on pattern matching. Here the abnormal system behaviour is defined at first by collecting the patterns of attack, and then define any other behaviour, as normal behaviour by matching them against the already recorded attacks.

### Anomaly Detection

Anomaly based intrusion detection system that offers classification of data into two types – normal data or threatened data. Based on this classification and observation of system activities, it detects intrusions and computer attacks. In this approach, normal data is recorded as baseline and threats are checked against this baseline.

## 2. PRIMARY APPROACH

The implementation of model will be checked against the standard dataset NSL-KDD [4]. This data set consists of selected records of the complete KDD'99 [5] data set, which is a complete data set and has been the most riotously used data set used for the study and assessment anomaly based intrusion detection. KDD'99[5] is the most preferred data set for intrusion detection as compared to other available data set because it is well labelled and contain several attack types and shows the multiple attack scenarios, whereas the other data set are limited. The NSL-KDD [4] data set that has been used for intrusion detection is refined data set of original KDD'99[5] data set [8]. KDD'99[5] data set was having the problem of duplicate or redundant data that was removed in NSL-KDD [4] data set. Redundant data is having the disadvantage of biasing the learning algorithms that is removed in NSL-KDD [8] data set, this make data set more realistic for attack detection.

The data set consist of 41 features, each representing the basic properties of network, these properties in combination helps to identify the attacks over the network. Data set contains 41 features and 1 class labelled as Normal and Anomaly, as the data is already labelled the success rate of implemented model can be calculated.

The approaches to detect the intrusions can be divided mainly into two main categories: Anomaly Detection and Misuse Detection. The Data set contains mainly 4 types of attacks "DOS, PROBE, U2R and R2L [8]" that is further divided into large number of attacks. A separate set of train and test data set is used to perform the evaluation, train set contains 24 training attack types [8] and test contains additional 14 types of attacks [8]. A separate set of train and test data makes detection more close to the real world attacks.

## 3. RELATED RESEARCH

A number of researchers have contributed a lot towards building strong Intrusion Detection system that raise new challenges and lay foundation for current work.

Learning the behaviour of attack indicators from the data set and then detecting the intrusions [6]. The technique is primarily used in order to detect unknown attacks that are occurring for the first time. It is a two step process:
Step 1: Based on the comparison of patterns, intrusion detection is done using equality matching algorithm.
Step 2: The ANN machine learning model based on back propagation is then implemented in order to detect the intrusion if still any aspect left.
The bottleneck in [16] is that at first step, patterns of old attacks need to be learnt before detecting any intrusion activity which is then tested by a single algorithm that can be done with many other learning algorithms as well.

Lisong Pei, Jakob Schütte, Carlos Simon in 2007-10-07 explained that there are two basic complementary trends in intrusion detection knowledge based, In knowledge base the knowledge about the attacks are taken for the detection of attacks[17].

The Lightweight Network Intrusion Detection System, LNID, is proposed system for intrusion detection. Here, a filtering scheme is proposed which consist of two filtering process; Tcpdump Filter and LNID Filter. Tcpdump Filter initiates packet filtering process using tcpdump tool that extracts TCP packets to telnet servers of internal LANs [3]. In [9], the authors used 10-fold cross validation and various other classifiers for detection of attacks. The classifiers were evaluated towards the prediction of classifiers by using 10-fold cross validation and since the learning stages and testing can be implemented on known attacks only hence the process is considered to find bottleneck in finding accuracy towards their detection.

In [11], an extended security technique for intrusion detection is proposed that claims cleaning of data in huge databases. It focuses on matching anomalous information with database policies that means most favourable situations to work in this technique are for known attacks with predefined policies. IP Flow-Based Intrusion Detection [12] approach finds the contents of attack by monitoring each and every packet however, inspection of packets is tough to perform at high speed. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani [8], prepared a new NSL-KDD [4] data set this was the refined version of old KDDCUP'99 [5] data set. They used different machine learning algorithms on their new data set. A separate set of training and test set was made, which make the detection more accurate for new unseen attacks. In [8] authors also mentioned that J48 that is ID 4 algorithm is most successful in detection of attacks with the accuracy of 81.05% .

Another author [10] also used the NSL-KDD[4] data set they preprocessed the data by feature selection and used different machine learning algorithms for the analysis and improved the detection rate on reduced features. The paper also mention that SimpleCart is the best algorithm for intrusion detection with the detection rate of 82.32% after feature selection, and is more successful algorithm to detect new and unseen attacks.

Mostly the authors had used KDDCUP'99[4] data set which is most widely used standard set, but this data set suffers from limitations due to duplication, which leads to the biasing in detection of attacks which are more frequent in data set like DOS and PROBE attacks. Some researchers had used NSL-KDD [4] data set which is improved version of original KDD'99 data set but all the experiments are done only on anomaly detection model.

Researchers tried to overcome the limitation of both anomaly detection and misuse detection by predicting the Hybrid models. Kim, G., Lee, S., & Kim, S presented the novel hybrid intrusion detection model using both the approaches of misuse and anomaly detection using C4.5 and One-class SVM but still hybrid model [13]contains the limitation of learning from the data set of known attack types due to which accuracy rate is only high only for the already learned attacks. This paper led our basis for the assumption that a hybrid model for detection of Intrusion can be built on selected features and implementing machine learning algorithms where no previous learning will be required.

## 4. PROPOSED IDS USING HYBRID MACHINE LEARNING ALGORITHM

The literature survey shows that very few authors' tried to improve the model for intrusion detection based on misuse detection concept. In misuse detection the prediction of attacks are done on the basis of the behaviour of the basic features of network, not by only matching the signature of attacks.

Some researchers implemented the hybrid approach between the anomaly and misuse detection, but still there model requires previous learning and model lack in detection of novel attacks. To build the hybrid model mostly researchers tried to build model over the anomalous approach but the mentioned model is based over the misuse approach to detect the attacks.

The proposed model consists of selected network features, by observing the behavior of attacks, analyzing the essential attributes of network packets and generating rules over them with hybrid of two machine learning algorithms IDS model is built where no previous learning is required for detection of attacks.

A brief introduction of J48 and SimpleCart is given. These two algorithms are required to build the hybrid model.

### 4.1. J48 Algorithm:

J48 that is also termed as C4.5 algorithm; here gain ratio is used to generate a decision tree that is actually a non binary tree in which the data sets are spilt with respect to the values of root node and the value of root node is based upon the features having highest value. Each and every node separately calculates its gain value and the process of calculation are carried until the process of prediction is carried.

J48 is basically a tree classifier of Weka Tool [14] developed by Quinlan [15] and is a version of C4.5 algorithm. In J48 Decision tree classifier, for attribute evaluation a decision tree is created

initially based on the attribute values obtained by training of data and the classification of data instances can be done on the basis of attributes influence by trained test data.

### 4.2. Classification & Regression Tree:

Classification & regression trees (CART) algorithm is a binary recursive partitioning technique that acts as predicts data algorithm. It decision trees are build using learning samples that are actually data from history with certain pre assigned classes. The cost of the decision tree so obtained is pruned by cost complexity pruning and the splits of the decision tree are selected using gini index [17]. Here, data is bifurcated into two subsets in such a way that each subset contains more homogeneous records than the previous subsets. The process is repeated recursively until either the homogeneity is met or some other stopping criteria is met and so each of the subset further splits into new subsets.

CART enables users to provide prior probability distribution [18] and has the ability to generate the regression trees. CART is performing the classification, based on decision trees. For constructing the tree, CART uses the training data in which classes are already assigned with the labels.

Rules over the network features are generated from the algorithms over the NSL-KDD [8] data set and implemented in the framed model.

## 5. PROPOSED HYBRID ALGORITHM

In the proposed hybrid algorithm first we will use the NSL-KDD data set; this data set will be processed in WEKA tool over machine learning algorithms.

1. Select training and test data (NSL-KDD data)
2. Generating rules over features of network
3. Building frame and making new rules by analyzing the generated rules.
4. Apply KDD data set and test the model.
5. Check evidence of attack, Calculate False positive and Negative

From processing different rules are produced by analyzing generated rules, network features of data set and results rules will be built for the network features.

A Framework will be deployed over these rules; this model will be tested against the KDD data set which is already labeled as attack and normal data to check the accuracy of the model.Fig.1 representing the Implementation setup model for the framework.

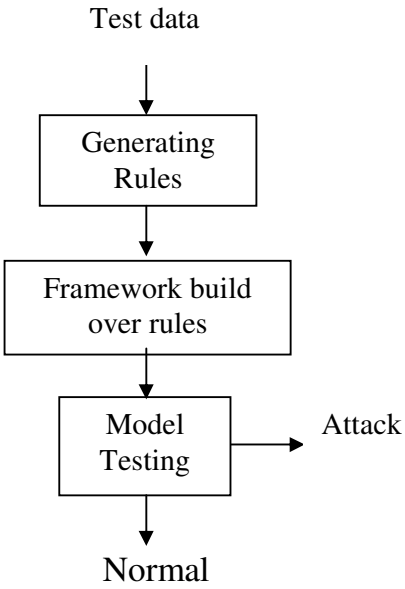


Figure1. Implementation Setup Model

Figure 1. is representing the use of test data to generate the rules over selective machine learning methods and then building the framework over the rules, this model will be tested on already verified data to predict the accuracy.

## 6. EXPERIMENTATION RESULTS AND PERFORMANCE COMPARISON

The rules built over the basic features of network are framed into the model and checked against the standard NSL-KDD data set for the prediction of attacks. All the models represented in table 1. works on large number of parameters to detect the intrusions but the proposed hybrid model contains only 29 features to detect the attacks which is far less than the available models and total number of features in the data set. Model also represents that high accuracy in attack prediction can be achieved on small number of features.

This model works on the misuse detection concept for intrusion. Model is capable of detecting the intrusions on the basis of behaviour of the basic features of network without any previous learning. It can be said that this model contains the advantage of machine learning techniques with misuse detection.

Table 1 is representing the results of proposed Hybrid Algorithm and the comparison chart with some previous work done by the researchers, Only [13] produces the better result but only for known attacks and still machine Learning is required but in proposed hybrid approach no learning is required and all the data was unknown for the model and no previous training previous is required.

Table 1: Detection Accuracy

| Classifier Algorithms & Proposed Algorithm | Detection Accuracy (%) | Detection Accuracy (%) |
|---|---|---|
| Proposed Hybrid Intrusion Detection Method | 88.23% | |
| Intrusion Detection Using Dimension Reduction [10] | 81.9375% using J48 | 82.32355% SimpleCart |
| Hybrid Anomaly & Misuse Detection[13] | 99.10% For Known Attacks | 30.05% For Unknown Attacks |
| Light Weight Network Intrusion Detection System[3] | 72.70% | |

### 6.1. Feature Selection in Misuse Detection Concept

This entire model presented in Figure 2 is based on anomaly detection approach where the machine learns the normal behaviour of network and when the test data is given, that is checked against the learned normal behaviour and predicted as normal or anomalous data. Based on feature selection and by analyzing the basic behaviour of system attacks a new model is developed. This model works on the misuse detection concept for intrusion, and is capable of detecting the intrusions on the basis of behaviour of the basic features of network without any previous learning. It can be said that this model contains the advantage of feature selection and machine learning techniques with misuse detection.

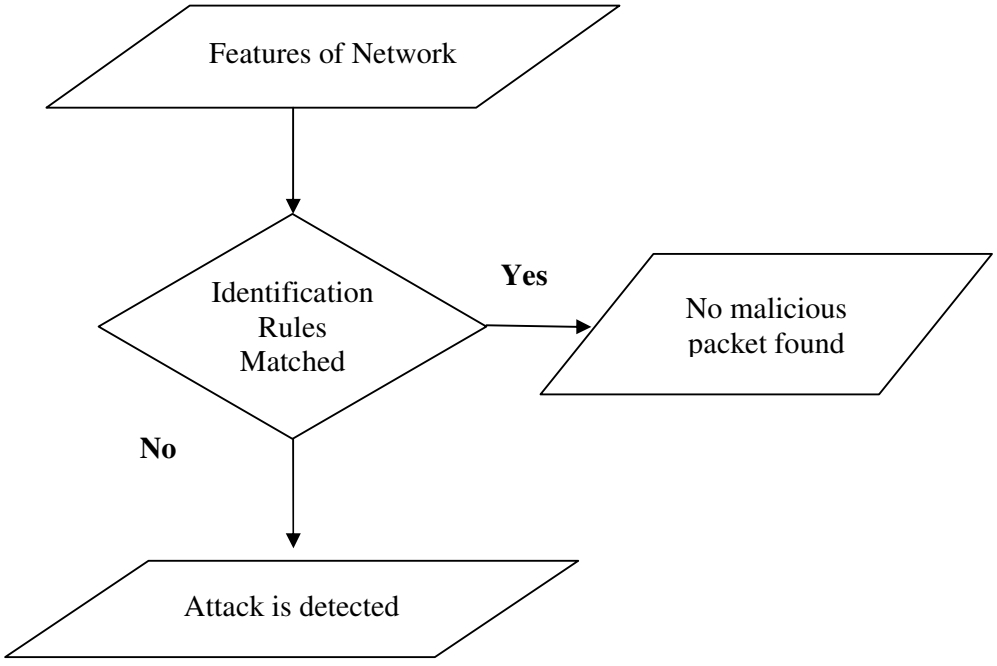


Figure 2. Misuse Detection Model

### 6.2. Result Analysis

The model presented in figure 3 is tested over more than 3000 data values of NSL-KDD data set which was randomly taken from the data set, this model gave the accuracy of 88.23% where all the attacks was novel for the model.

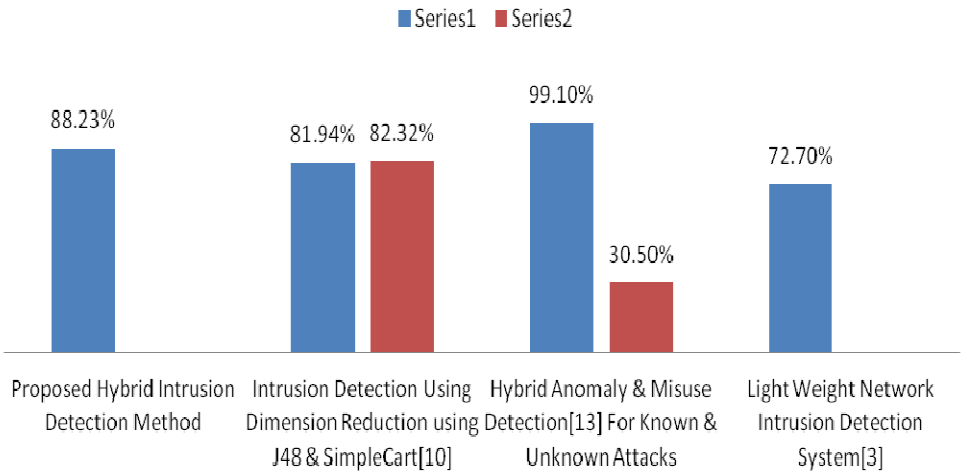


Figure 3. Comparative Analysis In Terms of Detection Accuracy

In Figure 3 bar graph is representing the detection accuracy 88.23 percent for the new hybrid model where no training is required for intrusion detection and other bars are representing the comparison with the existing models, in mostly all the models previous training is required and model are successful only on already known attacks.

## 7. CONCLUSION & FUTURE WORK

The proposed hybrid model for intrusion detection system assumes that higher number of features in the network needs not to be considered to achieve high accuracy. So, the proposed model is implemented over 29 features with the success rate of 88.23%.

Model for intrusion detection shows that by analyzing the basic behavior of network data; based on prior features, and by hybrid model the machine learning algorithms intrusion detection can be improved.

In the future work more parameters can be set for network features to improve the rate of intrusion detection, by further applying more techniques on proposed model the performance for IDS can be improved.

## REFERENCES

[1] Kumar, Sandeep, and Eugene H. Spafford. "A pattern matching model for misuse intrusion detection." (1994).

[2] Sinclair, C., Pierce, L., & Matzner, S. (1999). An application of machine learning to network intrusion detection. In Computer Security Applications Conference, 1999.(ACSAC'99) Proceedings. 15th Annual (pp. 371-377). IEEE.

[3] Chia-Mei Chen, Ya-Lin Chen, Hsiao-Chung Lin (2010) "An efficient network intrusion detection", Elsevier,vol 33 (4), pp. 477- 484.
[4] "Nsl-kdd data set for network-based intrusion detection systems. Available on: http://nsl.cs.unb.ca/NSL-KDD/ ,March 2009.
[5] KDD Cup (1999). Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[6] Anup K. Ghosh, Aaron Schwartzbard & Michael Schatz,(1999)" Workshop on Intrusion Detection and Network Monitoring" Santa Clara, California, USA.
[7] Lisong Pei, Schütte,J. (2007, July) Intrusion detection system, Carlos Simon.
[8] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani (2009) "A Detailed Analysis of the KDD CUP 99 Data Set" IEEE Symposium on computational intelligence in security and defence application.
[9] G.Meera Gandhi, Kumaravel Appavoo ,S.K. Srivatsa (2010) "Effective Network Intrusion Detection using Classifiers Decision Trees and Decision rules" Int. J. Advanced Networking and Applications, vol. 2(3), pp.686.
[10] Bajaj, K., & Arora, A. (2013). Dimension Reduction in Intrusion Detection Features Using Discriminative Machine Learning Approach. International Journal of Computer Science Issues (IJCSI), 10(4).
[11] Aarthy.R and P.Marikkannu (2012) "Extended security for intrusion detection system using data cleaning in large database" International Journal of Communications and Engineering, vol. 2(2),pp.56-60.
[12] Anna Sperotto, Gregor Schaffrath, Ramin Sadre, Cristian Morariu, Aiko Pras and Burkhard Stiller (2010) "An Overview of IP Flow-Based Intrusion Detection" IEEE communications surveys & tutorials, vol. 12(3): pp. 343.
[13] 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.
[14] "Waikato environment for knowledge analysis (weka) version 3.6.9. and 3.7.9"Available on :http://www.cs.waikato.ac.nz/ml/weka/
[15] Quinlan, J.: C4.5: Programs for Machine Learning, Publisher Morgan Kaufmann, San Mateo (1993)
[16] Khan, I,Q, (2009) Simultaneous prediction of symptom severity and cause in data from a test battery for Parkinson patients, using machine learning methods (Doctoral dissertation, Dalarna University).
[17] Rajput, S., & Arora, A. (2013). Designing Spam Model-Classification Analysis using Decision Trees. International Journal of Computer Applications, 75(10), 6-12.
[18] Lior Rokach and Oded Maimon, "DECISION TREES," Department of Industrial Engineering, Tel-Aviv University, Chapter-9,pp.181