Journal of Advances in Science and Technology
Vol. 15, Issue No. 1, March-2018, ISSN 2230-9659

Novel Data Mining Based Algorithms for Intrusion Detection in Communication Networks

R. N. Phursule[1]* Dr. Y. P. Singh[2]

[1] Ph.D. Scholar, Computer Science and Engineering Department, Kalinga University, Raipur, Chhattisgarh, India

[2] Research Guide, Computer Science and Engineering Department, Kalinga University, Raipur, Chhattisgarh, India

Abstract – There is a tremendous growth in the field of information technology due to which, network security is also facing significant challenges. The traditional Intrusion Detection System (IDS) is not able to prevent the recent attacks and malwares. Hence, Intrusion Detection System (IDS) which is an essential component of the network needs to be protected. IDS methodologies which are currently in use require human involvement to create attack signatures or to generating productive models for normal behaviour. In order to supply a potential another to expensive human input, we are in need of learning algorithms. The predominant task of such learning algorithm is to discover appropriate behaviour of IDS as normal and abnormal (system is under attack). The algorithm should be accurate and it should process the information in quick successions which is one of the major drawbacks in IDS because of the large amount of features. The intrusion detection plays an essential role in computer security. Data mining introduce to the process of separate hidden, previously unknown and useful information from huge databases. To detect patterns in the data set and use these patterns to find future intrusions data mining techniques help. Data Mining based Intrusion Detection System is combined with Multi-Agent System to improve the performance of the IDS. In the current era, there is ample knowledge in using Internet in social networks (such as instant messaging, video conferencing, etc.), the field of healthcare, various areas related to electronic commerce, banking, and services several other fields. As computer systems based on the network plays an ever more important in recent period once they have become the target of our criminals and enemies. Accordingly, we must determine the one of the best way to take our systems. The security of a computer system is compromised at the time of an intrusion occurs. Intrusion is nothing but the set of actions that intention is compromise the confidentiality, integrity or availability of a resource for example, illegally get super user privileges to attack and make out of the system (i.e, DOS), etc.

Keywords— IDS, Data mining, Classifier, Feature selection, Multi-Agent System.

----------♦----------

1. INTRODUCTION

Data mining is a procedure to extricate the data or learning consequently and astutely from an immense measure of information. Here during the time spent information mining, delicate data can be unveiled by trading off the person's entitlement to protection. Expanding interest of Privacy security in information mining gives me heading to explore about protection security information mining. Data mining is defines as the practice of examining large pre-existing databases in order to generate new information. The arrangement of illustrations used to take in the order model is known as the preparation dataset. Assignments identified with characterization incorporate relapse, which assembles a model from preparing information to foresee numerical values, and bunching, which bunches case to frame classifications. Characterization has a place with the class of directed taking in recognized from unsupervised learning. In managed taking in, the preparation information comprises of sets of information, and craved yields, while in unsupervised realizing there is no priori yield (Pontarelli, et. al., 2013).

An intrusion detection system observe network traffic for suspicious activity and take precaution the system or network administrator in order to take evasive action. It has a very important position in the network information security and it is considered as the second security gate after firewall. In recent years, intrusion detection method and key technology has become one of research focus in network security field (Levin, 2000).

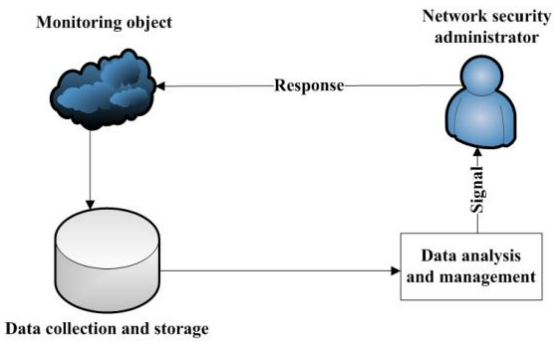


Figure 1. IDS structure

The intrusion detection does an important part in network security. Method of extricating invisible, formerly undisclosed as well as functional data from vast databases is mention to Data Mining. Hence data mining methods assist to recognize models in data set as well as utilize these models for identify destiny obtrusion. Data Mining based Intrusion Detection System is combined with Multi-Agent System to improve the performance of the IDS. In the current era, there is ample knowledge in using Internet in social networks (such as instant messaging, video conferencing, etc.), the field of healthcare, various areas related to electronic commerce, banking, and services several other fields (Kim & Park, 2003).

Data pre-processing is an information mining system that includes changing crude information into a justifiable arrangement. Certifiable information is regularly fragmented, conflicting, as well as ailing in specific practices or inclines, and is probably going to contain numerous mistakes. Information preprocessing is a demonstrated technique for settling such issues. Information preprocessing plans crude information for additionally processing (Chandrasekhar & Raghuveer, 2013). Data transferring by using the network protocols are TCP, UDP and HTTP. For KDD Cup 99 and to make a correlation with those frameworks that have been assessed on various sorts of attack we develop the five classes. One of these classes include simply the normal records and the other four hold diverse sorts of assaults (i.e., DoS, Probe, U2R, R2L), separately. Database standardization, or essentially standardization, is the way toward arranging the segments (qualities) and tables (relations) of a social database to lessen information excess and enhance information respectability (Mukkamala, et. al., 2005).

Feature extraction is an important part of an intrusion detection system. Feature selection is a technique for eliminating irrelevant and redundant features and selecting the most optimal subset of features that produce a better characterization of patterns belonging to different classes. Methods for feature selection are generally classified into filter and wrapper methods (Toosi & Kahani, 2007). Data cleansing or data cleaning is the way toward distinguishing and redressing (or expelling) degenerate or off base records from a record set, table, or database and alludes to recognizing inadequate, erroneous, wrong or unimportant parts of the information and afterward supplanting, changing, or erasing the grimy or coarse information (Ambusaidi, et. al., 2015).

Data integration includes consolidating in various sources and furnishing clients with a brought together perspective of them. This procedure ends up plainly critical in an assortment of circumstances, which incorporate both business, (for example, when two comparative organizations need to consolidate their databases) and logical (joining research comes about because of various bioinformatics vaults, for instance) spaces. Information reconciliation shows up with expanding recurrence as the volume and the need to share existing information detonates. It has turned into the concentration of broad hypothetical work, and various open issues stay unsolved (Amiri, et. al., 2011). Decentralization is the way toward appropriating or scattering capacities, forces, individuals or things far from a focal area or expert. While centralization, particularly in the administrative circle, is generally examined and rehearsed, there is no basic definition or comprehension of decentralization. The importance of decentralization may change to a limited extent on account of the distinctive ways it is connected. Ideas of decentralization have been connected to amass elements and administration science in private organizations and associations, political science, law and open organization, financial aspects and methodology (Kim, et. al., 2014).

Privacy-Preserving Data Mining (PPDM) is an information mining and factual databases inventive field where information digging calculations are broke down for reactions in information security. It is likewise called protection upgraded/security delicate information mining managing getting legitimate information mining comes about without picking up basic information esteems. This uncovers what number of various strategies and methods can be utilized as a part of a PPDM setting from a specialized point of view.

PPDM has risen to ensure the protection of delicate information and furthermore give substantial data mining come results. Figure 1.2 shows a distributed PPDM scenario which can achieve reasonable privacy and good accuracy. Regularly an exchange off amongst protection and exactness are should be made. From one perspective, security requires that the first information records must be completely jumbled before information mining investigation. Then again, precision needs that the "examples" in the first information ought to be mined out regardless of the

irritation data should be mined out in spite of the perturbation (Likun Liu et al 2012).

There are two major methods in PPDM

- First by using cryptographic representation.

- The other is by using heuristic algorithms which ensures that sensitive data is not revealed.

In this research paper, we used the several numbers of data classification techniques thus are, SVM (support vector machine), FFNN (Feed Forward neural network) and navies' bayes. The main objective of this paper is Classification algorithms are increasingly utilized for problem solving. In this study, efficiency of the various classification algorithms (such as k-NN, RBF, MLP, SVM) is compared with the proposed classification algorithms. The proposed classifiers perform comparative cross validation for existing classifiers. There are number of effective objective of the study.

2. RELATED WORKS

S. Pontarelli (2013)

In this paper [1], author represent that the help various, traffic-aware, modular method in the scheme of FPGA-based NIDS. Alternatively, traffic across equal modules, we categories and region homogeneous traffic, and celerity it to variously able hardware blocks, each helping a rule set tailored to the particular traffic classification. We developments and prove our method using the rule set of good Snort NIDS, and we analytically innovation the appear trade-offs and benefits, displaying resource storing up to eighteen percent depends on actual-world traffic statistics assembled from an operator backbone.

I. Levin (2000)

In [2], Kernel Miner is a recent data-mining instrument depends on constructing the optimal decision prediction. The instrument won second position in the KDD'99 Classifier Learning competition, August 1999. Also author explain the Kernel Miner's model and function used to discovering the competition task. The outcomes are received are examines and describe.

D. S. Kim (2003)

In this research paper [3], represent that approach of supplies for the SVM (support vector machine) to SVM IDS (based Intrusion Detection System). The SVM is a learning approach which is applied for the multiple numbers of projects. The Intrusion detection can be examines as the classification of the multi-class problem. The author utilized dataset from the 1999 KDD intrusion detection competition. The SVM IDS learning algorithm was analysis with the training data sets and check with the test datasets to discover the SVM IDS performance to the assaults. Author also analyses the significant of each feature to increase the entire performance of IDS. The outcomes of the research exhibits that enable SVM in IDS can be effective way for recognize the intrusions.

A. Chandrasekhar (2013)

In this research [4], the author innovate a recent approach by utilizing data mining approach thus are, radial basis and neuro-fuzzy SVM for the IDS. The innovation approach for the data mining has main 4 phase in which, the initial phase is consist the operating the FCM (Fuzzy C-means clustering). The second phase performs the neuro-fuzzy is trained, thus datasets indicates is trained with corresponding to neuro-fuzzy classifier associated with the cluster. Afterward, a vector for SVM separation is integrated and in pervious step, the dataset classification by using the radial SVM is action to recognize the intrusion has obtained or not. The datasets utilized KDD cup 99 dataset and they are the responsiveness, particularity and corrects as the analysis metrics arguments. The author approach is obtained good correctness for entire kinds of intrusions in network. It obtained about 98.94 percent correctness in DOS attack and getting of the 97.11 percent correctness in PROBE attack. In the R2L and U2R attacks, it obtained the correctness is 97.78 and 97.80 respectfully. They also differentiate the innovation approach with existing approach. These differentiate showed that the, ours approach is more effective than existing system.

S.Mukkamala (2005)

In [5], this research paper addresses using an object technique of various techniques. These techniques are hard and soft computing used for IDS. Due to rising the attacks of cyber, constructing effective IDS are necessary for preserve information system security, and although it residue an ambiguous objectives and issue. We analysis that, the performance of ANNs (Artificial Neural Networks), SVMs and MARS (Multivariate Adaptive Regression Splines). The author represented that objects of MARS, ANNs and support vector machine are higher to separates technique for IDS in concepts of classification correctness.

A. N. Toosi (2012)

In [6] this paper represented that, the main objective of this research paper is to integrated few soft computing approaches in to the classifying system to examines and categories exception from general aspects depends on the assault types in the computer network. In this Innovation work, the author investigates the some computer networks algorithm thus are, genetic algorithms (GA), fuzzy inference (FI), soft computing system and neuro-fuzzy network.

The neuro-fuzzy classifiers utilized to do primary classification of datasets. The fuzzy anticipation model depends on the outcomes of the neuro-fuzzy classifiers; generating final decision. Ultimately, in order to obtained the efficient output, the GA optimize the scheme of ours fuzzy decision engine.

A. M. Ambusaidi (201)5

In [7] paper, they introduce a new machine-learning-based data classification algorithm that is applied to network intrusion detection. The basic task is to classify network activities (in the network log as connection records) as normal or abnormal while minimizing misclassification. Although different classification models have been developed for network intrusion detection, each of them has its strengths and weaknesses, including the most commonly applied Support Vector Machine (SVM) method and the Clustering based on Self-Organized Ant Colony Network (CSOACN). Our new approach combines the SVM method with CSOACNs to take the advantages of both while avoiding their weaknesses. Our algorithm is implemented and evaluated using a standard benchmark KDD99 data set.

F. Amiri (2015)

In this research paper, they innovates two highlight choice calculations and look at the execution of utilizing these calculation separates to a common data based component choice approach. These feature selection algorithm necessary they employ of feature effectiveness determine. The author shown that, innovated by using the non-linear and linear determine-linear correlation coefficient and mutual information, for the feature selection. Ahead, we enhance IDS that used an expanding machine learning based approach, Least Squares Support Vector Machine. Reasonable work on KDD Cup 99 datasets address that our advancement shared data-based feature selection approach outcomes in recognize intrusions with most correctness, particularly for U2R (user to remote) and R2L (remote to login) attacks.

G. Kim (2014)

In [9] this paper, the author suggests that, present hybrid intrusion detection algorithm that hierarchically integrated a detection technique in a decay scheme is investigated. The first, abuse detection technique is created for based on the C4.5 decision tree algorithm and then the simple training data is share in to subsets by using this technique. The second technique, multiple one-class SVM technique is created for the decay subsets. The output of the second detection technique does not utilize the known attack data obliquely, but also creates the profiles of general properties exactly. The innovation hybrid IDS was analysis by organizes practical with the NSL_KDD datasets, which is a changed part, is called as KDD Cup 99 data set. The practical output shown that the investigation function is better than the conventional technique in theory of the detection rate for entire unknown and known attack, so these are maintain a positive rate. The author also proposed the innovation function beneficially decreases the large time ambiguity of the training and testing procedure.

R. Chitrakar (2014)

In [10] this research paper innovate the Half-partition strategy of chosen and managing non-support vector of the recent improvement of classification called as CSV (Candidate Support Vectors)-which are impartial to becomes SVM in the afterward enhancement of classification. The research work developed the CSV based Incremental SVM (CSV-ISVM) model; these are used to improve the SVM classification of the datasets. In this innovation work investigates the modifications to existing concentric-ring model and constrained group strategy. The efficiency of the innovation model compare with the ISVM model.

3. METHODOLOGY

In this chapter represents the different base classifier and how they work also present proposed ensemble classifier .The purpose of the present research is to study the classifier based text approaches for data mining methods. The researcher will identify techniques that were developed. Hence the purpose of this methodology is illuminating the concept of classifier based text approaches for data mining methods. This methodology will cover title of the study, significance of the study, aims and objectives of the study, research hypothesis and research design. This research has designed based upon descriptive study as it aims an in-depth analysis on classifier based text approaches for data mining methods.

The purpose of intrusion identification structures (IDS) is to defend from the signs and symptoms of protection problems. However, given that identification generally relies upon on the monitored information and has to calculate an intruder, the running of IDS involves threaten users' privacy. In this thesis, we suggest a brand new private saving method in intrusion calculation gadget via making use of cryptographic techniques to log files. It is able to meet the improved security of customers in addition to the security of network providers without TTP. To provide security we use cryptography approach the use of below encryption and decryption algorithms.

Algorithm 1: Encryption and Decryption

Encryption

Cipher(byte in[256], byte out[256], key_array round_key[Nr+1])

begin

State = in;

AddRoundKey (state, round_key[0]);

for i = 1 to Nr-1 stepsize 1 do

SubBytes (state);

ShiftRows (state);

MixColumns (state);

AddRoundKey (state, round_key[i]);

end for

SubBytes (state);

ShiftRows (state);

AddRoundKey (state, round_key[Nr]);

end

Decryption

Inv Cipher (byte in [265], byte out [265], key [256]) begin

State = in Add Round Key (state, Nr]) for round = Nr-1 step -1 down to 1

Add Round Key (state, key [Nr])

Mix Columns (state)

Shift Rows (state)

Sub Bytes (state)

end for

Shift Rows (state)

Sub Bytes(state)

Add Round Key (state, key[0, Nr-1])

out = state

end

Algorithm 2 Intrusion detection based on LS-SVM

Input: LS-SVM Normal Classifier, selected features (normal class), an observed data item x

Output: Lx - the classification label of x

begin

Lx ←classification of x with LS-SVM of Normal class

if Lx == “Normal” then

Return LX

else

do: Run Algorithm 3 to determine the class of attack

end

end

Algorithm 3 Attack classification based on LS-SVM using Brute Force Attack

Input: LS-SVM Normal Classifier, selected features, an observed data item x

Output: Lx - the classification label of x which attack found

Begin

Step1. c← first(x with normal classifier)

Step2. While c ≠ normal do all steps

Step3. Lx ←classification of x with LS-SVM of DoS class

Step4. If Lx==“DoS” then Return LX and Stop else goto

Step5. Lx ←classification of x with LS-SVM of Probe class

Step6. If Lx == “Probe” then Return LX and Stop else goto

Step7. Lx ←classification of x with LS-SVM of R2L class

Step8. If Lx == “R2L” then Return LX and Stop else goto 9

Step9.Lx == “U2R”; Return LX

Step10.if c ≠ normal goto Step2

Step11. Stop.

4. RESULTS AND DISCUSSION

The performance of the classifier has been firstly evaluated using three different dataset method and algorithm runs on each dataset. The experimental results of the classification algorithms with using feature selection method and ensemble three classifiers for a KDD datasets are presented below graphs.

Accuracy= (TP+TN / TP + TN + FP + FN) * 100

Below figure 2 is showing the comparative result for accuracy.

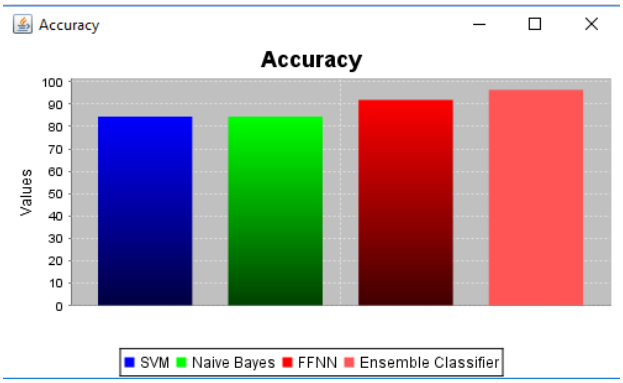


Figure 3: Accuracy rates of ensemble classifiers

Here we compare the accuracy value of SVM, Naïve Bayes, FFNN and Ensemble classifier using KDD dataset. SVM accuracy is 84%, Naïve bayes is also 85%, FFNN is 92% and Ensemble classifier has 97%.

Below figure 3 is showing the comparative result for Precision rates of ensemble classifiers and precision formula:

$$\text{Precision}=\frac{TP}{TP+FP}$$

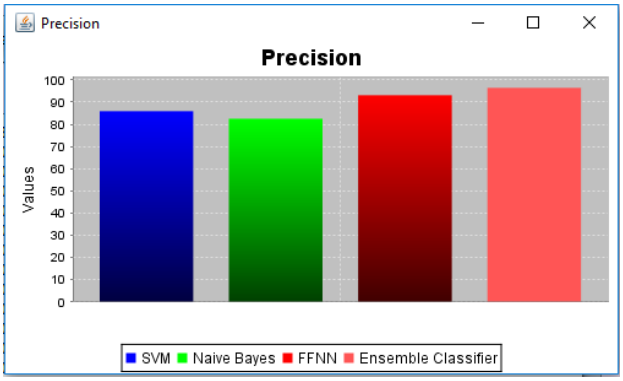


Figure 3: Precision rates of ensemble classifiers

Here we compare the accuracy value of SVM, Naïve Bayes, FFNN and Ensemble classifier using KDD dataset. SVM accuracy is 85%, Naïve bayes is also 82%, FFNN is 90% and Ensemble classifier has 92%.

Below figure 4 is showing the comparative result Recall rates of ensemble classifiers and recall formula:

$$\text{Recall}=\frac{TP}{TP+FN}$$

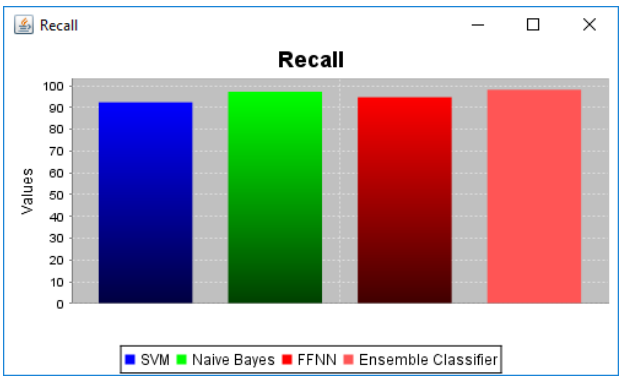
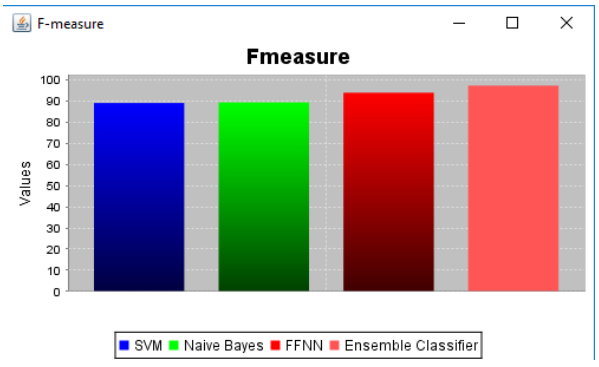


Figure 4: Recall rates of ensemble classifiers

Here we compare the recall value of SVM, Naïve Bayes, FFNN and Ensemble classifier using KDD dataset. SVM measure is 92%, Naïve bayes is also 97%, FFNN is 93% and Ensemble classifier has 98%.

Below figure 5 is showing the comparative result F-measure rates of ensemble classifiers.

Here we compare the f-measure value of SVM, Naïve Bayes, FFNN and Ensemble classifier using KDD dataset. SVM f-measure is 90%, Naïve bayes is also 90%, FFNN is 94% and Ensemble classifier has 98%.



5. CONCLUSION

This paper major objective of this research is to provide a promising solution to address the intrusion detection problem. Traditional methodologies such as machine learning methods, classifications and cluster

methods, data mining techniques used for network security and it's provide the better efficiency. This research work focuses on classification algorithms and feature selection methods to increase the identification performance and to reduce the time required to carry out the computations for intrusion detection systems. This work dealt with the problem of feature selection, which is of great importance in intrusion detection due to high dimensional data. To improve the accuracy rate of the detection system, the classifiers are hybridized and evaluated on the benchmark intrusion detection dataset, KDDCup'99 from UCI machine learning repository. Data mining methods and classification approaches have been applied for intrusion detection system to differentiate normal and abnormal behaviour. The proposed feature selection methods, namely: Flexible mutual information based feature selection (FMIFC) and hybrid feature selection algorithm (HFS) was evaluated on standard data mining classification algorithms: Naïve Bayes (NB), Decision Tree and Support Vector Machine (SVM). The effectiveness of feature selection methods was tested using hybrid approach based on Artificial Neural Network algorithms. In the proposed method, the three ensembles in feature ranking methods are fusion, selection and hybrid methods. The effectiveness of ensemble classifier is tested through an all three basic classification algorithms. The proposed hybrid classifier produces best results using the features of hybrid methods. Also, the performance of proposed method is compared with traditional classifiers: Naïve Bayes (NB), Support Vector Machine (SVM), and Feed Forward Neural Network. Future work can be extended using various bio-inspired algorithms for feature selection and classification with real-time network datasets. The privacy preserving Online Analytical Processing (OLAP) can be integrated with the proposed framework to enhance and improve the effectiveness and the flexibility of the IDS system.

REFERENCES

A. Chandrasekhar, K. Raghuveer (2013). An effective technique for intrusion detection using neuro-fuzzy and radial svm classifier, in: Computer Networks & Communications (NetCom), Vol. 131, Springer, pp. 499–507.

A. M. Ambusaidi, X. He, P. Nanda (2015). Unsupervised feature selection method for intrusion detection system, in: International Conference on Trust, Security and Privacy in Computing and Communications, IEEE, 2015.

A. N. Toosi, M. Kahani (2007). A new approach to intrusion detection based on an evolutionary soft computing model using neuro fuzzy classifiers, Computer communications 30 (10) pp. 2201– 2212.

D. S. Kim, J. S. Park (2003). Network-based intrusion detection with support vector machines, in: Information Networking, Vol. 2662, Springer, pp. 747–756.

F. Amiri, M. RezaeiYousefi, C. Lucas, A. Shakery, N. Yazdani (2011). Mutual information-based feature selection for intrusion detection systems, Journal of Network and Computer Applications 34 (4) pp. 1184–1199

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

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

I. Levin (2000). Kdd-99 classifier learning contest: Llsoft's results overview, SIGKDD explorations 1 (2) pp. 67–75.

M. Tavallaee, E. Bagheri, W. Lu, A.-A. Ghorbani (2009). A detailed analysis of the kdd cup 99 data set, in: Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications, pp. 1–6.

R. Chitrakar, C. Huang (2014). Selection of candidate support vectors in incremental svm for network intrusion detection, Computers & Security 45, pp. 231–241.

S. Mukkamala, A.H. Sung, A. Abraham (2005). Intrusion detection using an ensemble of intelligent paradigms, Journal of network and computer applications 28 (2) pp. 167–182.

S. Pontarelli, G. Bianchi, S. Teofili (2013). Traffic-aware design of a high speed fpga network intrusion detection system, Computers, IEEE Transactions on 62 (11) pp. 2322–2334.

Corresponding Author

R. N. Phursule*

Ph.D. Scholar, Computer Science and Engineering Department, Kalinga University, Raipur, Chhattisgarh, India