Intrusion detection technique using Coarse Gaussian SVM

Bhoopesh Singh Bhati*

Department of Computer Science and Engineering,
Ambedkar Institute of Advanced Communication,
Technologies and Research,
Geeta Colony, Delhi, India
Email: bhoopesh.cse@gmail.com
*Corresponding author

C.S. Rai

USIC&T,
Guru Gobind Singh Indraprastha University,
Dwarka, New Delhi, India
Email: csrai_ipu@yahoo.com

Abstract: In the new era of internet technology, everybody is transferring data from place to place through the internet. As internet technology is improving, different types of attacks have also increased. To detect the attacks it is important to protect transmitted information. The role of Intrusion Detection System (IDS) is very imperative to detect various types of attacks. Although researchers have proposed numerous theories and methods in the area of IDS, the research in area of intrusion detection is still going on. In this paper, Coarse Gaussian Support Vector Machine (CGSVM) based intrusion detection technique is proposed. The proposed method has four major steps namely, Data Collection, Pre-processing and Studying data, Training and Testing using CGSVM, and Decisions. In implementation, KDDcup99 data sets are used as a benchmark and MATLAB programming environment is used. The results of the simulation are presented by Receiver Operating Characteristics (ROC) and Confusion Matrix. The proposed method achieved detection rates as high 99.99%, 99.95%, 99.53%, 99.19%, 90.57% for DOS, Normal, Probe, R 2 L, U 2 R respectively.

Keywords: information security; intrusion detection; machine learning; Coarse Gaussian SVM; anomaly detection; networks security.

Reference to this paper should be made as follows: Bhati, B.S. and Rai, C.S. (2021) 'Intrusion detection technique using Coarse Gaussian SVM', *Int. J. Grid and Utility Computing*, Vol. 12, No. 1, pp.27–32.

Biographical notes: Bhoopesh Singh Bhati is pursuing his PhD degree from the G.G.S.I.P. University Delhi. He obtained his MTech degree in Information Security and BTech degree in Computer Science and Engineering from the G.G.S.I.P. University Delhi. He is working as an Assistant Professor in the Department of Computer Science and Engineering of Ambedkar Institute of Advanced Communication Technologies and Research, Govt. of NCT, Delhi. He has published various research papers in international journals and conferences. His current research area is information security.

C.S. Rai is a Professor with the University School of Information and Communication Technology since 2011. He obtained his ME degree in Computer Engineering from SGS Institute of Technology and Science, Indore in 1994 and completed his PhD degree in the area of Neural Network from Guru Gobind Singh Indraprastha University in 2003. He has many publications in international/national journals and conferences. He was conferred with the Best Teacher Award of the university for the academic year 2007–2008. His teaching and research interests include: artificial neural systems, computer networks and signal processing.

1 Introduction

Nowadays, the internet has become one of the most important needs of every individual. The internet is the global system for inter-communication between computer networks or a system which uses the network protocol (TCP/IP). Owing the recent development, our day to day life fully depends on internet network technology. It is very important to assure reliable operation on the network (Lazarevic, 2005). The number of attacks on the network is increasing day by day. In order to enhance the system security or network security, intrusion detection is used (Liao et al., 2013). So this becomes the important reason for the development of Intrusion Detection Systems (IDS).An IDS has the ability to monitor different types of activities on the system. An attacker is the person or the organiser which performs the doubtful activities like destroy, expose, disable, alter or gain in order to access the unauthorised data of a computer network. The main purpose of IDS is to detect the attack and report it, so that the administrator becomes careful and can takes steps to guard against these attacks (Gupta et al., 2014).

Intrusion Detection System observes the network traffic on a system. It alerts the system when some doubtful activity is occurring. The main purpose of IDS is to give information to people employed in IT that network intrusion has occurred. IDS technology is one of the crucial parts of different working areas and IS widely used to detect, identify and respond to an abnormal activity that occurs in a system (Bhati and Rai, 2016). Different types of network attacks such as Denial of Service (DOS), Probe, Remote to User (R2L) attack, and User to Root (U2R) can be detected by IDS. The purpose of IDS is to collect data from a computer and examine this data in order to find safety episodic events and then present the consequence to the administrator (Tavallaee et al., 2009) Intrusion detection techniques can be broadly categorised into following types:

Signature-based intrusion detection: Signature IDS is also known as misuse IDS and sometimes known as knowledge-based intrusion detection. By using some form of pattern matching, misuse detection is done. String matching is the simplest form of pattern matching (Lundin and Jonsson, 2002).

When intrusion occurs, the signature of the intrusion which is the unique bit pattern can be observed in the network. Whenever the intrusion occurs, it runs the program of pattern matching in order to match the unknown network activity with the signature of all known intrusion. The network connection is classified as intrusion if a signature match is found. Misuse intrusion detection detects all the intrusions whose signatures are available in the database. The drawback of misuse IDS is that it is not able to detect intrusions whose signatures are not available in the database.

Anomaly detection: Anomaly detection is also known as the behaviour-based intrusion detection models. In this type of intrusion detection, deviation in the behaviour of an occurring event is observed. If deviation is great then the event is treated as intrusion. The anomaly detection is used to detect recent attacks that cannot be overcome by existing patterns. The main advantage of anomaly detection is that it can detect the unknown attacks. The drawback of anomaly detection is that the probability of false positives is relatively high because it is difficult to proceed with the correct set of network activity parameters to make profiles (Ramadas et al., 2003).

2 Related work

Support Vector Machine (SVM) is used to classify two classes of data. Support vector machine is built in order to find the optimal hyperplane that maximise the margin between two classes of data using gradient descent. SVM is useful for solving the binary classification problem and used to classify the attacks. In SVM, kernel function plays a vital role. Whenever, the SVM is not capable to separate two classes of data then this problem can be solved using kernel function (Sherif and Ayers, 2003).

A linear algorithm is mapped into non-linear classification by using SVM. A non-linear classification can be efficiently performed by using the kernel function. Kernel function includes radial bias function, polynomial, Coarse Gaussian etc. which can be used to divide two classes of data by constructing a decision line. Decision line is used to separate both the classes of data. That decision line is called the hyperplane. A hyperplane is a linear decision surface which is used to split two classes of data into two parts. A hyperplane is R^n in an n–1 dimensional subspace.

SVM is a new form of supervised machine learning algorithm. Classification and regression problem can be solved by using SVM. It is mostly used for classification problem. SVM can be used for intrusion detection. The data points nearest to the hyperplane are called the support vector, because it helps for the creation for the hyperplane.

Mukkamala et al. (2002) proposed a scheme for intrusion detection using SVM and neural network. In this research, the KDD data sets have been used. Their paper is also concerned for audit trial reduction. The output of SVM and neural network is measure and compared. This paper presented by Mohammed and Sulaiman (2012) an improved model is given for Intrusion Detection. In this paper authors used real-time data sets. Mulay et al. (2010) proposed new multiclass SVM for intrusion detection and used KDDcup99 data sets. Chen et al. (2005) proposed a scheme for application of SVM and ANN for intrusion detection. In this research, they used two data mining techniques one is Support Vector Machine (SVM) and another one is Artificial Neural Network (ANN). For increasing the potential of intrusion detection they use two schemes, one is simple frequency-based scheme and other is $tf \times idf$ scheme. By using SVM with $tf \times idf$ scheme, they achieve a better result as compared to ANN with simple frequency-based scheme. The SVM with simple frequency-based encoding method gives 10.00% false positive rate whereas ANN with simple frequency-based encoding method

gives 40.72% false positive rate. Authors use SVM with *tf* × *idf* encoding scheme for attack detection which gives 8.53% false positive rate whereas ANN with *tf* × *idf* encoding scheme gives 39.23% of false positive rate. Frequency-based encoding method is used rather than a sequence-based encoding method.

In the paper presented by Bo and Yuan (2009) a superior method is proposed, which overcomes the demerits of statistical-based detection technique. In the research of Bo and Yuan (2009) real-time data set is used to check the performance of proposed IDS technique. This real-time data set has normal records and malicious records. Deng et al. (2003) proposed a hierarchical and distributed IDS for wireless networks. In simulation DOS attack data sets was used. The results were analysed based on detection rate and false positive rate. Their proposed IDS was network independent. Khan et al. (2007) proposed a clustering tree-based SVM technique for IDS. Clustering tree is used to reduce training set. DAPRA1998 data set is applied for the experiments. Results have been analysed in terms of accuracy, false positive rate and false negative rate. Kim and Park (2003) proposed a network-based IDS using SVM. For implementation KDDcup99 data set has been used and the results have been analysed the basis of accuracy. Kang et al. (2012) proposed an anomaly intrusion detection approach based on one-class classification technique. In this research, authors used DAPRA1999 data sets and provided better results.

Wang et al. (2009) proposed a scheme for Intrusion Detection based on Particle Swarm Optimisation – Support Vector Machine (PSO-SVM). KDDCup99 data set is used and PSO-SVM model is applied on intrusion detection problem. Implementation of PSO is simple and it is an optimisation technique. They analysed that PSO-SVM gives high-detection rate as compared to SVM algorithm. They proposed a hybrid PSO which is based on feature selection algorithm in order to build IDS.

3 Proposed method

Our proposed method has four major steps namely data collection, pre-processing and studying the data sets, training and testing and decisions (Figure 2). Explanation and working is given below step by step.

Figure 1 Support vector machine

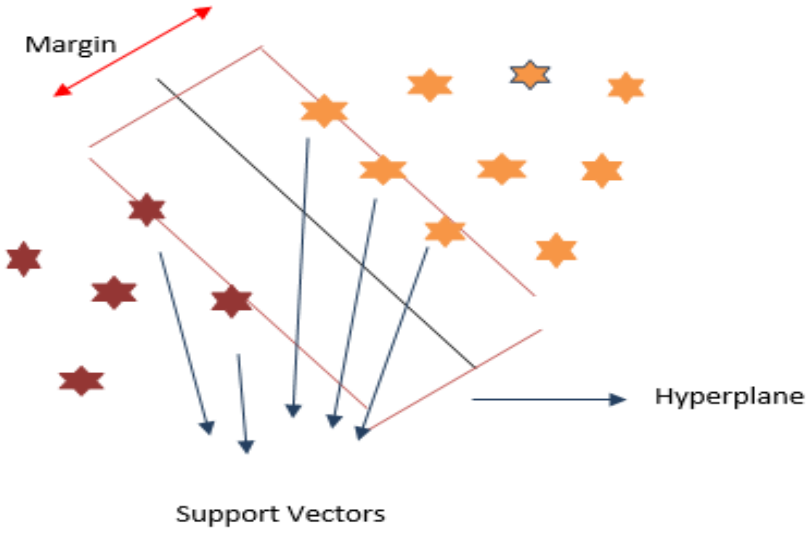
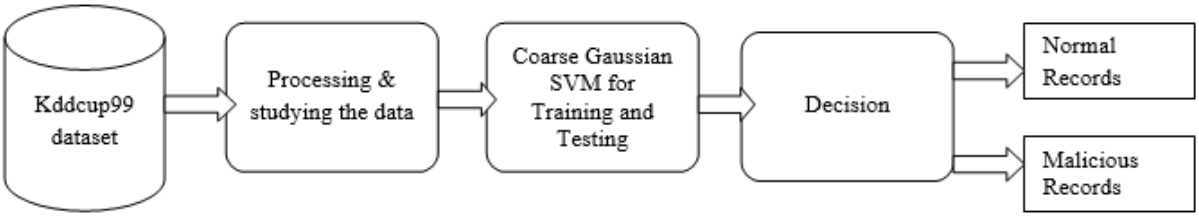


Figure 2 Proposed method



3.1 Data collection

In this proposed scheme, KDDcup99 data set is used. This data set contains huge amount of connection records approximately five million records (Sahu et al., 2014). The KDD99 data set is large in size therefore KDD 10% is used as a data set in our implementation. In KDD 10%, 34 records are numerical data type and seven records are character data type. It is a mixture of normal records and attacks. The attacks are divided into four categories (Siddiqui and Naahid, 2013).

Denial of service (DOS): To handle legitimate request, attackers tries to do computing and make the memory resources busy. DOS attack is further divided into two types one is crashing service and other one is flooding service. The examples of DOS attack are ping of death, back, Neptune, smurf, etc.

Remote to local (R2L): In this type of attack, attackers tries to access the user password by doing some computing with security e.g., guess_passwd, ftp_write, spy, phf, etc.

User 2 root (U2R): In this type of attack, attacker tries to access normal user account e.g., eject, buffer_overflow, loadmodule, etc.

Probing: In this type of attack, attackers want to gain the information of user machine e.g., satan, nmap, mscan, etc.

3.2 Pre-processing and studying data

Pre-processing is a technique used to transform the raw data into desirable format. Pre-processing and studying data step is used to remove conflict in a data set during training time. It is very important process in data mining. Pre-processing includes cleaning the data set, normalisation, feature selection, feature extraction, transformation. The two methods of pre-processing which are feature extraction and feature selection impact directly on the accuracy of the model. In pre-processing phase, information should not be lost from the data set used in implementation. Doing pre-processing, detection rate is improved that filters the false rates. Pre-processing is also used to discover the attack pattern and make the data set efficient (Revathi and Malathi, 2013).

3.3 Training and testing

Training and testing step is a crucial step of proposed method. After processing, this step comes in to picture. Here, Coarse Gaussian SVM is used to train the model and test them. CGSVM is a non-linear learning approach. It makes coarse distinctions between classes. The main application areas of CGSVM are regression learning and classification.

Suppose P is number of predictors then CGSVM can be mathematically defined as (Farayola et al., 2018):

$$\text{Gaussian kernel: } K\left(X, X_i\right) = e^{-Y\left|X - X_i\right|^2} \tag{1}$$

$$\text{Kernel scale set} = \left(P\right) * 4 \tag{2}$$

The configuration which is used in implementation of proposed method is given in Table 1. Kernel function is used to solve non-linearity problem. Box constraint level is responsible to hold valid value of multiplies. Multiclass method is responsible to overcome the multiclass problem. In this implementation, five-fold testing cross validation is used. That means, four folds are used in training and one fold is used for testing.

Table 1 Configuration of proposed method

Kernel	*Coarse Gaussian*
Box constraint level	1
Kernel scale mode	Manual
Manual kernel scale	25
Multiclass method	One versus One

3.4 Decisions

In decision phase, proposed method is capable for taking decisions in an efficient way by recording or detecting all the normal and malicious activity. Proposed method detects all types of attack categories and gives detection rate for all attack categories i.e. DOS, U2R, R2L and Probe.

4 Empirical evaluation and results

Proposed method is implemented on MATLAB. Results have been analysed through ROC and Confusion Matrix. ROC is also called performance curve (Kumar and Selvakumar, 2011). Here, for every class of attack, ROC is drawn. The angle between true positive rate and false positive rate should be 90° in ROC plot for a perfect intrusion detection technique. With the help of area under curve, detection accuracy is observed. Confusion Matrix is a pictorial way to represent the outcome of proposed method. It shows the result in term of true class and predict class. The terminology which is used in ROC and Confusion Matrix is discussed below:

True positive: when IDS is capable of identifying correct network and malicious activity.

True negative: attack does not take place in a system and also no alarm is produced by IDS.

False positive: IDS give false alarm rate when no attack has occurred.

False negative: when attack take place in a system but IDs does not generate alarm.

$$\text{True Positive Rate (sensitivity)} = \frac{\text{true positive}}{\textit{true positive} + \textit{false positive}} \tag{3}$$

$$\text{False Positive Rate (1-specificity)} = \frac{\text{false positive}}{\textit{false positive} + \textit{true Negative}} \tag{4}$$

Figures 3, 4, 5, 6 and 7 show the ROC graphs using the proposed method for DOS, Normal, Probe, R2L, and U2R respectively. Figure 8 shows the overall Confusion Matrix. The outcome of this proposed method is demonstrated by all mentioned figures. The detection rate is given in Table 2.

Figure 3 ROC of DOS

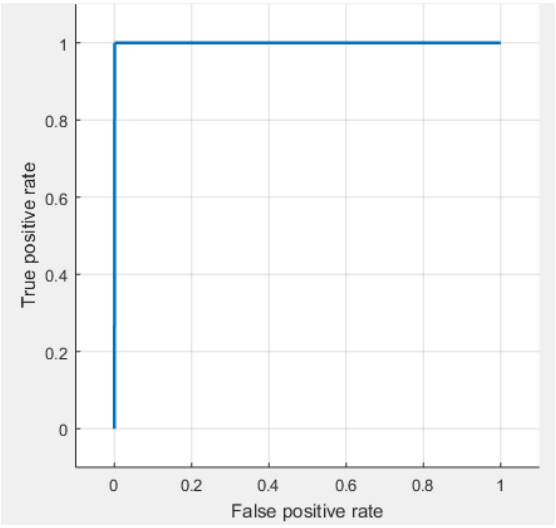


Figure 4 ROC of normal

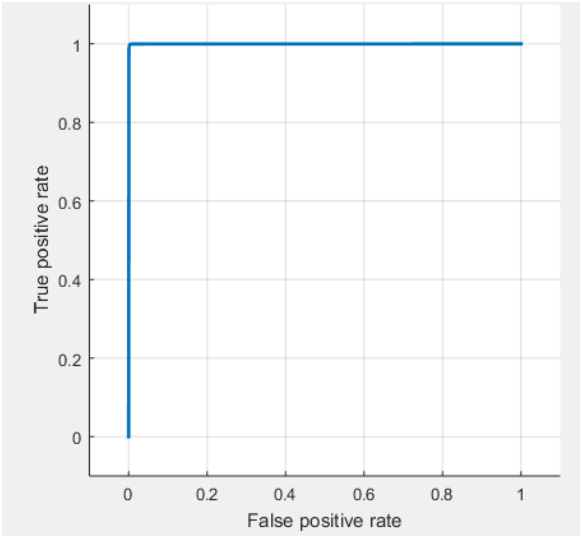


Figure 5 ROC of probe

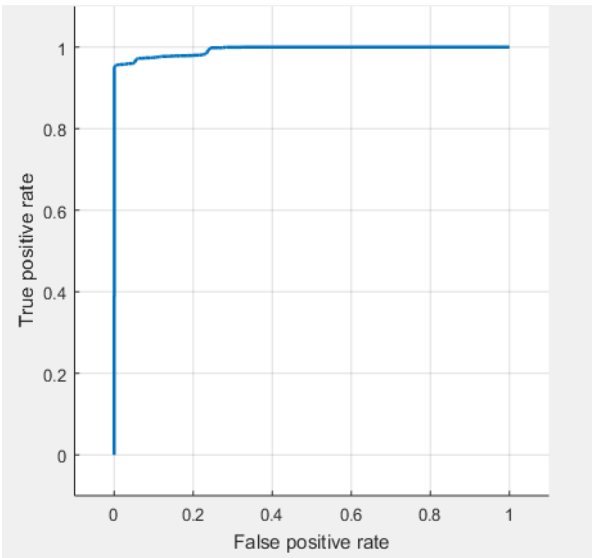


Figure 6 ROC of R 2 L

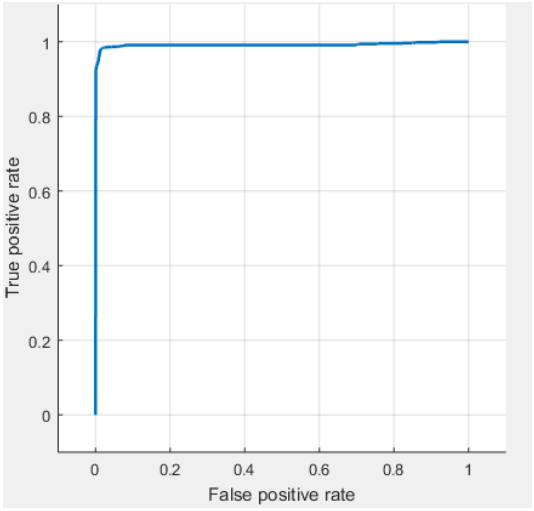


Figure 7 ROC of U 2 R

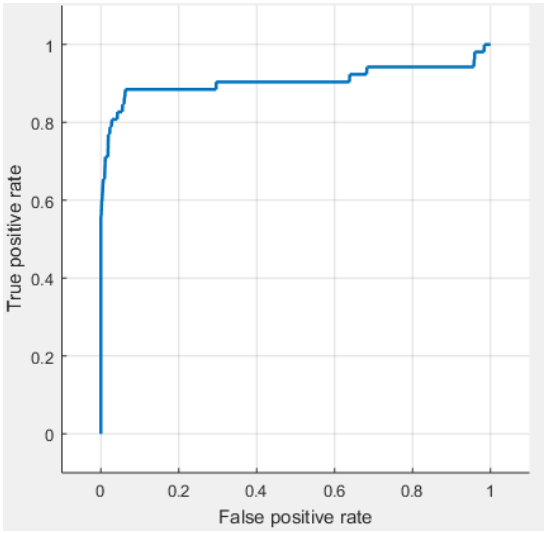


Figure 8 Overall confusion matrix

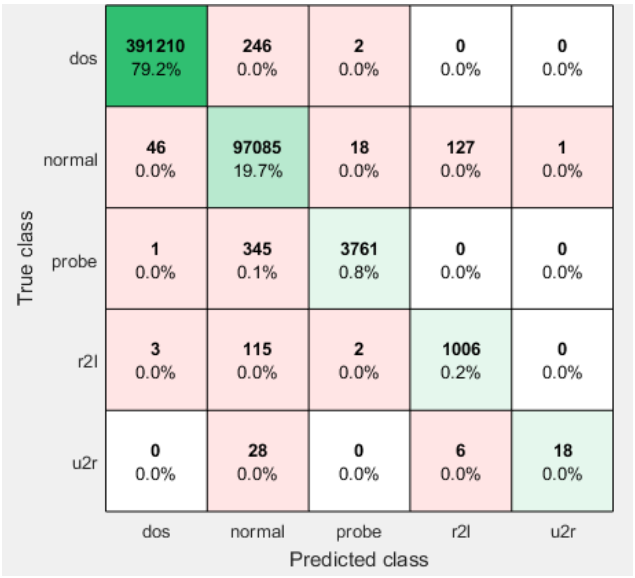


Table 2 Detection rate

Type of attack class	*Detection rate (%)*
DOS	99.99
Normal	99.95
Probe	99.34
R 2 L	99.19
U 2 R	90.57

5 Conclusion

In this paper, Coarse Gaussian SVM-based intrusion detection has been proposed and implemented. KDDcup99 data set and MATLAB have been used in implementation. Results have been analysed through ROC and Confusion Matrix. Accuracy of different attacks class is observed with the help of ROC. Simulated results shows that proposed method provides high detection rates of 99.99%, 99.95%, 99.53%, 99.19%, 90.57% for DOS, Normal, Probe, R 2 L, U 2 R, respectively. In future, the proposed method may be applied on real-time data sets and optimisation technique may be involved in CGSVM in order to obtain more accurate intrusion detection techniques.

References

Bhati, B.S. and Rai, C.S. (2016) 'Intrusion detection systems and techniques: a review', *International Journal of Critical Computer-Based Systems*, Vol. 6, No. 3, pp.173–190.

Bo, L. and Yuan, C.Y. (2009) 'The research of intrusion detection based on support vector machine', *Proceedings of the International Conference on Computer and Communications Security*, IEEE, pp.21–23.

Chen, W.H., Hsu, S. H. and Shen, H.P. (2005) 'Application of SVM and ANN for intrusion detection', *Computers and Operations Research*, Vol. 32, No. 10, pp.2617–2634.

Deng, H., Zeng, Q.A. and Agrawal, D.P. (2003) 'SVM-based intrusion detection system for wireless ad hoc networks', *Proceedings of the IEEE 58th Vehicular Technology Conference*, IEEE, Vol. 3, pp.2147–2151.

Farayola, A.M., Hasan, A.N. and Ali, A. (2018) 'Efficient photovoltaic MPPT system using Coarse Gaussian support vector machine and artificial neural network techniques', *International Journal of Innovative Computing Information and Control (IJICIC)*, Vol. 14, No. 1, pp.323–339.

Gupta, A., Bhati, B.S. and Jain, V. (2014) 'Artificial intrusion detection techniques: a survey', *International Journal of Computer Network and Information Security*, Vol. 6, No. 9, pp.51–57.

Kang, I., Jeong, M.K. and Kong, D. (2012) 'A differentiated one-class classification method with applications to intrusion detection', *Expert Systems with Applications*, Vol. 39, No. 4, pp.3899–3905.

Khan, L., Awad, M. andThuraisingham, B. (2007) 'A new intrusion detection system using support vector machines and hierarchical clustering', *The VLDB Journal*, Vol. 16, No. 4, pp.507–521.

Kim, D.S. and Park, J.S. (2003) 'Network-based intrusion detection with support vector machines', *International Conference on Information Networking*, Springer, Berlin, Heidelberg, pp.747–756.

Kumar, P.A.R. and Selvakumar, S. (2011) 'Distributed denial of service attack detection using an ensemble of neural classifier', *Computer Communications*, Vol. 34, No. 11, pp.1328–1341.

Lazarevic, A. (2005) *Managing cyber threats: issues, approaches, and challenges*, Springer Science+ Business Media, Incorporated.

Liao, H.J., Lin, C.H.R., Lin, Y.C. and Tung, K.Y. (2013) 'Intrusion detection system: a comprehensive review', *Journal of Network and Computer Applications*, Vol. 36, No. 1, pp.16–24.

Lundin, E. and Jonsson, E. (2002) *Survey of Intrusion Detection Research*, Chalmers University of Technology.

Mohammed, M.N. and Sulaiman, N. (2012) 'Intrusion detection system based on SVM for WLAN', *Procedia Technology*, Vol. 1, pp.313–317.

Mukkamala, S., Janoski, G. and Sung, A. (2002) 'Intrusion detection: support vector machines and neural networks', *Proceedings of the IEEE International Joint Conference on Neural Networks (ANNIE)*, IEEE, USA.

Mulay, S.A., Devale, P.R. and Garje, G.V. (2010) 'Intrusion detection system using support vector machine and decision tree', *International Journal of Computer Applications*, Vol. 3, No. 3, pp.40–43.

Ramadas, M., Ostermann, S. and Tjaden, B. (2003) 'Detecting anomalous network traffic with self-organizing maps', *International Workshop on Recent Advances in Intrusion Detection*, Springer, Berlin, Heidelberg, pp.36–54.

Revathi, S. and Malathi, A. (2013) 'Data preprocessing for intrusion detection system using swarm intelligence techniques', *International Journal of Computer Applications*, Vol. 75, No. 6, pp.22–27.

Sahu, S.K., Sarangi, S. and Jena, S.K. (2014) 'A detail analysis on intrusion detection datasets', *Proceedings of the IEEE International Advance Computing Conference (IACC)*, IEEE, pp.1348–1353.

Sherif, J.S. and Ayers, R. (2003) 'Intrusion detection: methods and systems: Part II', *Information Management and Computer Security*, Vol. 11, No. 5, pp.222–229.

Siddiqui, M.K. and Naahid, S. (2013) 'Analysis of KDD CUP 99 dataset using clustering based data mining', *International Journal of Database Theory and Application*, Vol. 6, No. 5, pp.23–34.

Tavallaee, M., Bagheri, E., Lu, W. and Ghorbani, A.A. (2009) 'A detailed analysis of the KDD CUP 99 data set', *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*, IEEE, pp.1–6.

Wang, J., Hong, X., Ren, R.R. and Li, T.H. (2009) 'A real-time intrusion detection system based on PSO-SVM', *Proceedings of the International Workshop on Information Security and Application (IWISA'09)*, Academy Publisher, pp.319–321.