# An Efficient Ensemble mechanism for Intrusion Detection

Dr. Pramod1,

1Associate Professor, Department of ISE,

PESITM,

Shimoga, India

Email: pramod74@pestrust.edu.in

# Dr. Sunitha B S 2

2Associate Professor, Department of CSE,

PESITM,

Shimoga, India

Email: sunitha.bs@pestrust.edu.in

Lohith C3

3Assistant Professor, Departmentof CSE,

Gopalan College of Engineering and Management,

Bangalore, India

Email: lohithchandrashekhar@gmail.com

# Sabin T T4

4Assistant Professor, Department of ISE,

SJCIT,

Chickballapur, India

Email: sabinsree@gmail.com

# **ABSTRACT**

Most classification and regression problems can be solved by training multiple learners, not by building a learner from the data. The boosting algorithm represents a recent advancement in classification methodology, which has the ability to enhance the performance of weak learners and transform them into strong ones. This technique operates by sequentially incorporating a classification algorithm into the updated weights of training samples, utilizing the majority voting technique of a sequence of classifiers. The AdaBoost algorithm is an efficient algorithm that combines weak algorithms to create a powerful classifier which can classify the training dataset with high accuracy. Based on the simulation outcomes, it is evident that the AdaBoost classifier can attain a remarkable level of search precision within a brief computational duration and at a minimal expense, in comparison to the classification approach. We have introduced a policy model to produce standard and assault categories and executed an online access mechanism to authorize or prohibit network access.

**Keywords**—Clustering, optimization, weak learners, strong learners, AdaBoost.IDS(Intrusion Detection System)genetic algorithm (GA), SVM (Support Vector Machine), Extreme Learning Machine (ELM)

## INTRODUCTION

With significant advancements in the digital governance of e-commerce, social media, etc., the Internet is playing an increasingly significant role on a global scale. However, fear, criminal activity, and cyber attacks that started to build and launch highly complex attacks driven by destructive intentions have made the internet vulnerable. Our network resources and gadgets need to be secure, i.e., private, intact, and available [1] [2]. The activity of identifying and collecting data pertaining to malicious actions aimed at compromising the integrity of computer and network security is commonly referred to as an intrusion detection system. [1]. It is a crucial and delicate component of the deep security system, which also consists of intrusion detection, firewalls, program wrappers, and scanning and patching for vulnerabilities, access control, and encryption. Security will be needed to protect the Internet's vital infrastructure.

The main advantage of intrusion detection is the training and installation effort with the inference engine. We have antivirus and detection systems implemented in our network and we constantly strive to develop and carry out new attacks. As soon as information about a new attack is gathered by detection systems, it must be quickly integrated with current detection systems in order to avert additional damage from the new attack as quickly as possible. However, due to training difficulties and large amounts of data, retraining models for existing and new attacks is often time-consuming. By the time new types of detection are available, new types of intruders may have done significant damage.

Intrusion is unauthorized access to hidden resources or restricted domains. This is how attackers gain unauthorized access to your network or private network. An attack is any suspicious or malicious activity on a network or computer. This is unauthorized interference with someone else's property. In retrospect, attackers try to identify security vulnerabilities before attacking her systems. To detect unwanted behavior that compromises security such as privacy, integrity, or availability. The remarkable progress in technology and the Internet has given rise to significant challenges in computer security. Contemporary advanced research endeavors to enhance diagnostics through various machine learning, data mining, and cognitive algorithms. These algorithms are classified into two detection methods: static detection (offline) and dynamic detection (online). The latter is a mechanism that promptly identifies suspicious activities on the network. Dynamic detection methods are more effective, reliable, and efficient than static methods.

**Learning Techniques**: The act of generating a model from data is commonly referred to as learning or training. There exist numerous learning methodologies, with the two primary approaches being supervised learning and unsupervised learning. In supervised learning, the objective is to accurately predict the target in real-time. Conversely, unsupervised learning does not depend on labeled information and endeavors to uncover distributional patterns within the data. In certain domains, a singular class or multiple coordinators may serve as the facilitator. Multiple taxonomies are also called hierarchical taxonomies. During model design, feature selection and training techniques reduce computational cost, model size, range, and accuracy. In hierarchical systems, planners are often seen as weak and incapable of planning effectively. The search accuracy is close to zero. However, in many layers weak class systems can be combined into strong layers and defined with good detection rates [4]. Numerous learning algorithms are presently accessible for the classification of samples or exemplars in datasets. Efficient algorithms encompass linear discriminant analysis, neural networks, decision tree Navie-Bayes classifiers, nearest neighbors, and SVM [5].

The ensemble classification method is simply the process of combining many weak learner methods to create an efficient learner that can effectively organize the learning process. Good and efficient predictions can be generated if the settings are close or correct to the actual values ​​[6]. Essentially, clustering algorithms are supervised algorithms because they can be trained to make their predictions true. You can find good ideas for making good predictions for a particular problem. The main idea of ​​clustering methods is to use one central learner to generate many ideas [7].

## LITERATURE SURVEY

The underlying philosophy of the classifier theorem posits that a particular classifier serves to offset the errors of another classifier. Nevertheless, the mere training of a classifier in isolation may prove insufficient in resolving the intended problem, given its lack of connectivity. The base manager represents a distribution employed in constructing a hierarchy. To build a classification system, we can consider a variety of weak learners such as support vector machines, nearest neighbors, and neural networks. Basic learning content is systematically created as extensions and containers. This upgrade adds weak learners and trains strong learners that provide better results and more accurate predictions. The boosting algorithm is founded upon the work of Kearns and Valiant [1989] and their two complexity classes, which address the fundamental inquiry of whether weak and hard learning problems are equivalent. This inquiry holds significant importance, as an affirmative response would enable any weak learner to transform into a strong learner. In practice, however, it is often easy to find weak learners, but difficult to find strong learners. Schapire [1990] showed the answer in the affirmative, leading to the development of boosting algorithms [8].

Due to the swift expansion of the Internet and the escalating worldwide availability of online content [16], the incidence of cybercrime has experienced a significant surge. Both end users and businesses are now vulnerable to cyber threats. It is important to put in place protective measures such as firewalls and IDSs. A firewall acts as an entry point, allowing or denying the passage of packets based on predefined criteria. In extreme cases, all network traffic may be blocked. Conversely, IDS automates the monitoring of computer networks. However, the continuous flow of data in such networks poses major challenges in developing effective IDSs, and a new approach utilizing online classification of datasets is proposed in this study to address this problem. Introduced in This method involves an incremental naive Bayesian classifier and uses active learning to achieve results on small sets of labeled data that are expensive to obtain. This approach involves his two sets of actions: offline, where data are preprocessed, and online, where the NADAL online method is introduced. A comparative analysis conducted on the NSL-KDD standard dataset has revealed several advantages of the proposed method. Firstly, it effectively overcomes the challenges associated with streaming data. Secondly, it mitigates the high cost associated with instance labeling. Thirdly, it demonstrates improved accuracy and kappa values when compared to the incremental naive Bayesian approach. Consequently, it can be inferred that this method is highly suitable for Intrusion Detection System (IDS) applications.

In our increasingly internet-reliant landscape, the main drawback of unauthorized intrusion into computer systems has escalated [17]. Intrusion refers to illicit access or activity within a computer tem. This highlights the growing importance of intrusion detection techniques to bolster overall computer system security. Intrusion detection involves identifying, preventing, or addressing intrusion attempts. This paper centers on an Intrusion Detection System driven by a GA. The technique applies GA to enhance network Intrusion Detection Systems (IDSs). It provides an overview of IDT, genetic algorithms and related detection methods. The paper delves into GA parameters, evolution processes, and their intricate details. Notably, this implementation uniquely considers both temporal and spatial attributes of network connections when encoding connection information into IDS rules, which aids in identifying complex anomalous behaviors. The focus of this work lies in TCP/IP network protocols. Intrusion Detection Systems can be classified intotwo groups based on their scope: Network-based IDS (NIDS) and Host-based IDS. NIDS observes intrusions by monitoring network traffic through devices like Network Interface Cards (NICs). Conversely, Host-based IDS monitors file and process activities within a specific host's software environment. Some host-based IDSs also analyze network traffic to detect attacks against a host.

A study [18] explored the impact of macro-level opportunity indicators on cyber theft victims and applied criminal opportunity theory to assess risk exposure. Use Internet access state patterns and other structural state characteristics to measure risk and estimate cyber damage impact. Moreover, there exists a positive correlation between the prevalence of cyber theft and the proportion of users who solely access the Internet from their residence. This study discusses the theoretical implications of these results. The role of link and node characteristics in social networks is investigated using the node classification method OS-ELM [19]. This method considers both node attributes and interactions for classification. Additionally, the ELM is used for intrusion detection within the IDS. Comparing the performance of SVM and ELM shows that the accuracy is comparable, but ELM has faster processing time. ELM is better than SVM in detecting intruders and the detection process takes less time.

## PROPOSED METHODOLOGY

An ensemble learning process trains a large number of learners rather than individuals and combines the end results of various learners to achieve better results than weak individual learners. Henceforth, it is commonly referred to as a multiple classification system [6]. A collective process always consists of combining several ideas to create the best idea. In other words, organization is the activity of gathering a large number of weak learners for the purpose of producing strong learners. A set of words is often reserved for how multiple ideas can be derived in a single base learner. Broadly speaking, many classification systems also cover different ideas that do not come from a single elementary school student. The illustration depicted in Figure 1 portrays a building block comprising n weak learners amalgamated into a single strong learner. These weak learners, also referred to as core learners, are required to be derived from a core learning algorithm, such as decision trees, neural networks, or any other form of learning algorithm. The primary objective of combinatorial techniques is to merge the predictions of numerous models generated through learning techniques to enhance the adaptability or resilience of a solitary model.



**Figure 1: Architecture of Simple ensemble approach**

There exist essentially two categories of learners, namely homogeneous learners and heterogeneous learners. The combinatorial method uses a special basic learning method that produces the same learners, i.e. learners who learn the same type, resulting in a unified system [8]. Different types of learners experience different processes and many learning algorithms are used. The generic skills of a particular group are often superior to the basic learner skills. Combinatorial techniques are of significant interest due to their ability to transform weak learners into strong learners. Weak learners possess the capability of generating random predictions, whereas strong learners exhibit the capacity to produce precise predictions.

1. **Proposed predictive model**

An example of our proposed prediction is shown in Figure 1. Details of the model training and testing process are shown in Figure 2. The process of training and testing comprises two distinct components, namely the online system and the offline system. In the offline method, the purpose of network traffic memory is to save training time, generate a function with the correct data structure, and match it with online network traffic behavior and corresponding model training. The proposed model comprises three primary components, namely data processing, inference engine and model training

**Figure 2: Model trained to classify traffic in network**

**Table 1: Confusion-Matrix for NSLKDD Dataset**

##

**Table 2: Confusion-Matrix for KDD Corrected Dataset**

****

**Table 3: Proposed model Performance Evaluation**



## RESULTS AND DISCUSSION



**Figure 3: Accuracy Estimation of Existing Methods**

Above graph depicts the Accuracy of Existing Algorithms and its comparison.



**Figure 4: Accuracy Estimation of Proposed Methods**

Above graph depicts the Accuracy of Proposed Algorithms and its comparison.

## CONCLUSION

Using clustering techniques to distinguish between regular traffic and congestion improves initial detection with minimal computation and cost compared with a single classifier. AdaBoost is a highly effective algorithm for reducing false alarms through the identification of false positives. Utilizing identical datasets for both training and testing the proposed model results in a significant increase in accuracy and a decrease in error.. However, the accuracy is relatively low when using different datasets for training and testing. Combining three weak layers such as SVMs, neural networks, and decision trees can outperform individual layers. It is concluded that the augmentation of learners in the combinatorial model enhances the precision of recognition and diminishes the likelihood of erroneous circumstances arising during each iteration.

## REFERENCES

1. Michael Kearns and Leslie Valiant. Cryptographic limitations on learning boolean formulae and finite automata. Journal of the ACM (JACM), 41(1):67–95, 1994.
2. William Stallings. Network and internetwork security: principles and practice, volume 1. Prentice Hall Englewood Cliffs, 1995.
3. Edward Amoroso. Intrusion detection: an introduction to internet surveillance, correlation, trace back, traps, and response. Intrusion. Net Book, 1999.
4. Bernhard Scholkopf and Alexander J Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.
5. Wray Buntine and Tim Niblett. A further comparison of splitting rules for decision-tree induction. Machine Learning, 8(1):75–85, 1992.
6. Robert E Schapire. The strength of weak learnability. Machine learning, 5(2):197–227, 1990.
7. Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1):119–139, 1997.
8. Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In Proceedings of the ninth annual conference on Computational learning theory, pages 325–332. ACM, 1996.
9. Erico N de Souza and Stan Matwin. Extending adaboost to iteratively vary its base classifiers. ´ In Advances in Artificial Intelligence, pages 384–389. Springer, 2011.
10. Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning, 36(1-2):105–139, 1999.
11. Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In Computational learning theory, pages 23–37. Springer, 1995.
12. Giorgio Valentini and Thomas G Dietterich. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. The Journal of Machine Learning Research, 5:725–775, 2004.
13. Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. The Journal of Machine Learning Research, 1:113–141, 2001.
14. Simon Haykin and Neural Network. A comprehensive foundation. Neural Networks, 2(2004), 2004.
15. David H Wolpert and William G Macready. No free lunch theorems for optimization. Evolutionary Computation, IEEE Transactions on, 1(1):67–82, 1997
16. P. Alaei and F. Noorbehbahani, “Incremental anomaly-based intrusion detection system using limited labeled data,” in *WebResearch (ICWR), 2017 3th International Conference on*, 2017, pp. 178–184
17. Aman V. Mankar , Tushar C. Ravekar, “A Study of Intrusion Detection System using Advanced Genetic Algorithm” International Research Journal of Computer Science (IRJCS) ISSN: 2393-9842 Issue 11, Volume 3 (November 2016)
18. H. Song, M. J. Lynch, and J. K. Cochran, “A macro-social exploratory analysis of the rate of interstate cyber-victimization,”*American Journal of Criminal Justice*, vol. 41, no. 3, pp. 583–601, 2016
19. Sun, Y., Yuan, Y., & Wang, G. (2015). An on-line sequential learning method in social networks for node classification. *Neurocomputing, 149*, 207–214. <http://doi.org/10.1016/j.neucom.2014.04.074>