**INTRUSION DETECTION SYSTEM USING**

**MACHINE LEARNING**

***Information Security Analysis and Audit (CSE3501)***

Review1 Report

By

**ARADHITA BHANDARI (18BIT0313)**

**ALGORITHM: AdaBoost Algorithm**

***TEAM MEMBERS:-***

|  |  |  |
| --- | --- | --- |
| **Team Member Name** | **Registration Number** | **Algorithm** |
| HITARTHI DARSHAN VORA | 18BIT0014 | XGBOOST |
| MADHUR BHATIA | 18BIT0054 | SVM |
| ARADHITA BHANDARI | 18BIT0313 | ADABOOST |
| AYUSHI UDASI | 18BIT0129 | LOGISTIC REGRESSION |

*Under the guidance of*

**Prof. Sumaiya Thaseen I – SITE**

Fall Semester 2020-21

SLOT-G1

****

**School of Information Technology & Engineering**

**1. Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. (Shahraki, A., Abbasi, M., & Haugen, Ø. ,2020)**

*i) Technique/algorithm used and why it was chosen (motivation)*

The paper aims to show the advantages of using boosting approaches in IDSs and to evaluate the effectiveness of well-known boosting approaches comparatively. Boosting was chosen over stacking and bagging techniques as it is the most powerful method that benefits from the ensemble learning idea. Ensemble learning is a subset of machine learning that is essentially using multiple models instead of using a single model so as to improve the performance in machine learning.

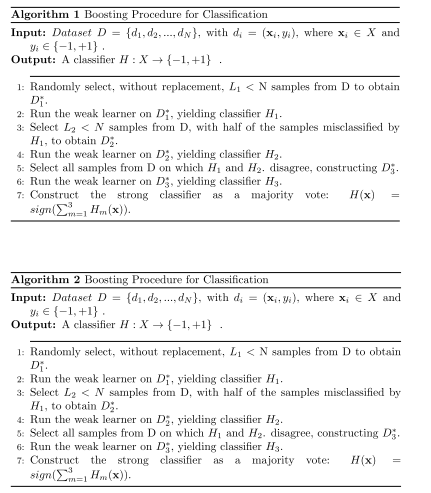
AdaBoost, short for Adaptive Boosting was formulated by Yoav Freund and Robert Schapire and is a machine learning meta-algorithm that is often referred to as the out-of-the-box classifiers. It can be used in conjunction with other types of learning algorithms to improve performance, as the output of the other algorithms is combined into a weighted sum that represents the final output of the boosted classifier. It is adaptive as the subsequent “weak learners: are tweaked in favour of the instances that were misclassified by previous classifiers.

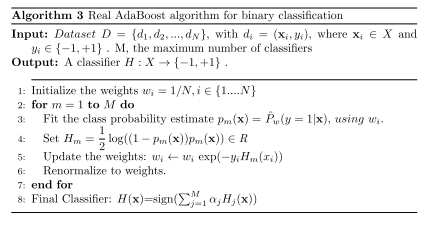
The paper evaluates three of the most power powerful boosting binary classifiers for accurate network intrusion detection based on comparative results. The algorithms used are Gentle AdaBoost, Modest AdaBoost and Real AdaBoost. The paper referred to previous papers in which Gentle AdaBoost had a better performance (Friedman et al., 2000) and in which Modest AdaBoost had better performance (Vezhnevets and Vezhnevets, 2005).

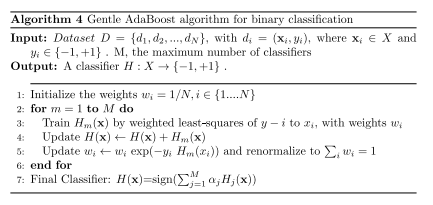
Major preprocessing steps were used before using the boosting algorithms to increase the classification accuracy and five famous intrusion detection datasets were used to evaluate the models.

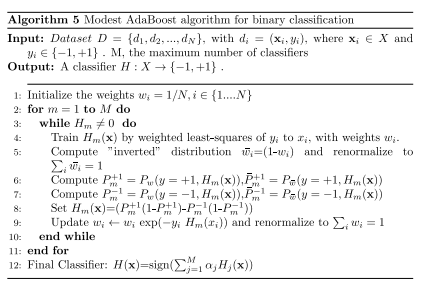
The models are compared based on different aspects such as running time, error rate and their performance on frequent retraining. The performance on frequent retraining is important as unlike most other ML applications, IDSs require frequent training based on various circumstances to remain efficient even as new attacks are developed and discovered.

*ii) Architecture/ model/pseudocode developed*

**







*iii) Datasets analyzed in the paper with the performance results*

The datasets used are

1. KDD Cup ‘99 dataset - k = 41; N = 5,000,000 Used - k = 34, N = 50,000

* Introduced for the Fifth International Conference on Knowledge Discovery and Data Mining through computer network simulations (Protić, 2018).
* Derived from 1998 DARPA
* Attacks - DoS, user to root, remote to local, and probe
* <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

1. UNSW-NB15 dataset - k = 49; N = 2,540,044 Used - k = 39, N = 36,628

* By the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) in 2015
* Attacks - Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms (Moustafa, Slay, J., 2015)
* <https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/>

1. TRAbID dataset - Used - k = 36, N = 36,628

* Viegas, et al. in 2017
* Attacks - DoS (HTTP flood, ICMP flood, SMTP flood, SYN flood, TCP keepalive), port scans (ACKScan, FIN-Scan, NULL-Scan, OS Fingerprinting, Service Fingerprinting, UDP-Scan, XMAS-Scan)
* <https://secplab.ppgia.pucpr.br/?q=trabid>

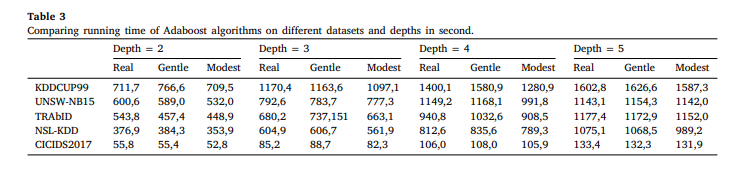
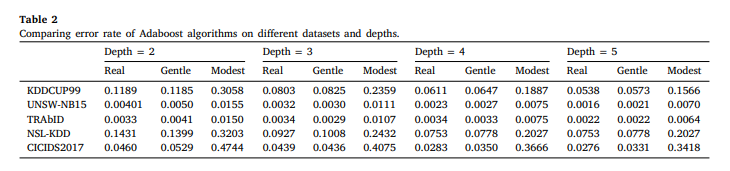
1. NSL-KDD dataset - k = 43, N = 125,973 Used - k = 23, N = 40,000

* Canadian Institute for Cybersecurity in 2009
* Improvement on the KDD CUP 99
* Attacks - DoS, privilege escalation (remote-to-local and user-to-root), probing
* <https://www.unb.ca/cic/datasets/nsl.html>

1. CICIDS2017 dataset - k = 80, N = 3,119,345 Used - k = 65, N = 3100

* Canadian Institute for Cybersecurity in 2017
* Captured- 9 a.m., Monday, July 3, 2017 to 5 p.m. Friday July 7, 2017
* Attacks - Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet and DDoS
* <https://www.unb.ca/cic/datasets/ids-2017.html>

Preprocessing for datasets : NULL value removal, transform of the nominal categorical features, feature standardization



*iv) Any comparison done with the previous techniques to specify that the proposed method is superior*

Comparing the error rates and running times of the evaluated AdaBoost algorithms shows the Gentle and Real AdaBoost have almost the same performance, their error rate curves are stable after a few numbers of iteration as it is an important factor in IDSs, because the ML models in IDSs needs to be retrain, Also, they have better performance than Modest AdaBoost as they improve the error rate about 70% but they are slower than Modest only about 7% in intrusion detection systems.

For the NSL-KDD subsets used, Real, Modest and Gentle AdaBoost achieved error rates of 0.0011, 0.0033, and 0.0011 respectively which can be compared to the lightweight IDS based on decision tree which gave a detection rate of 98.4% (Sindhu et al., 2012). The Ramp Loss K-SupportVector Classification-Regression model for multi-class intrusion detection by Bamakan et al. (2017) had a total detection rate of 98.86%. For TRAbID, a decision tree technique achieved 99.41% in DoS detection and 98.42% was achieved by Naive Bayes in Probing detection. Thus AdaBoost, especially Gentle and Real AdaBoost, are two algorithms that can be employed with incredibly high accuracy in IDSs.

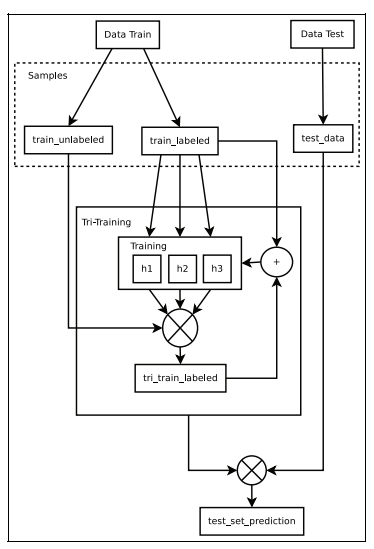
**2. Semi-Supervised tri-Adaboost algorithm for network intrusion detection. (Yuan, Y., Huo, L., Yuan, Y., Wang, Z., 2019)**

*i) Technique/algorithm used and why it was chosen (motivation)*

The algorithm proposed is a Semi-supervised tri-adaboost algorithm. It utilizes three different AdaBoost algorithms, namely the Discrete, Gentle and Real AdaBoost, as weak classifiers on both continuous and categorical data, constituting the decision stumps in the tri-training method. The chi-square method is also used to reduce the dimension of feature and improve computational efficiency.

Unsupervised learning is required for machine learning in IDSs as labelling data for learning purposes is a tedious manual task. However, unsupervised learning leads to a lower detection efficiency and accuracy. This algorithm was proposed to decrease the IDSs error rate and to find a model that is capable of incorporating new data with a good generalization capability. This is achieved with semi-supervised learning in a semi-supervised IDS by combining tri-training with three different AdaBoost algorithms. This allowed for a minimization of false positives, reduction of time consumption and increase in detection accuracy.

*ii) Architecture/ model/pseudocode developed*

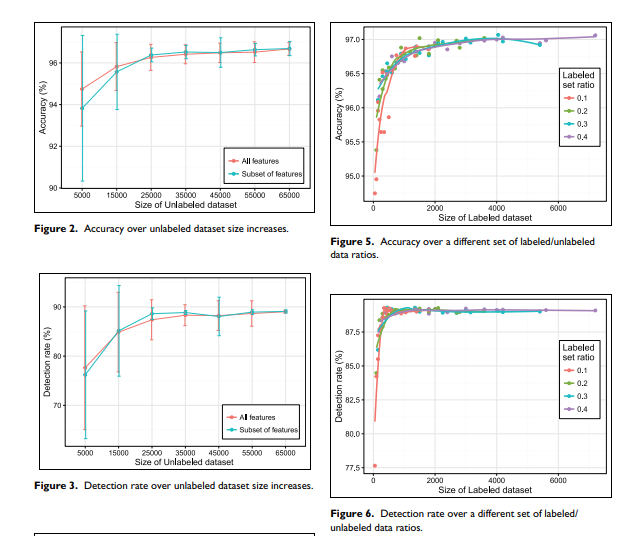


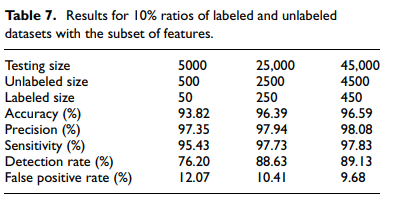
*iii) Datasets analyzed in the paper with the performance results*

Knowledge Discovery and Data Mining CUP 1999 dataset

* Attacks - Dos, PROBE, Remote to Local and User to Root

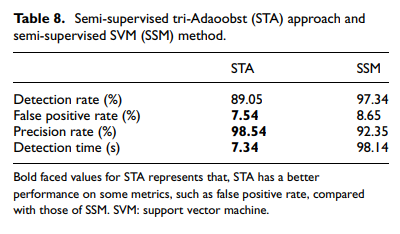
Preprocessing: Data is processing into labelled and unlabelled datasets and a separate dataset for testing





*iv) Any comparison done with the previous techniques to specify that the proposed method is superior*

Comparisons were made with a semi-supervised SVM (SMM) model proposed by Jimin et al. Specifically, the labeled dataset consists of 4000 samples, with 3000 normal and 1000 attacks. The unlabeled training dataset has 10,000 samples, with 7000 normal samples and 3000 attacks. Finally, the testing set consists of 30,000 samples, with 25,000 normal samples and 5000 attacks. Table 8 shows a basic comparison of our results.



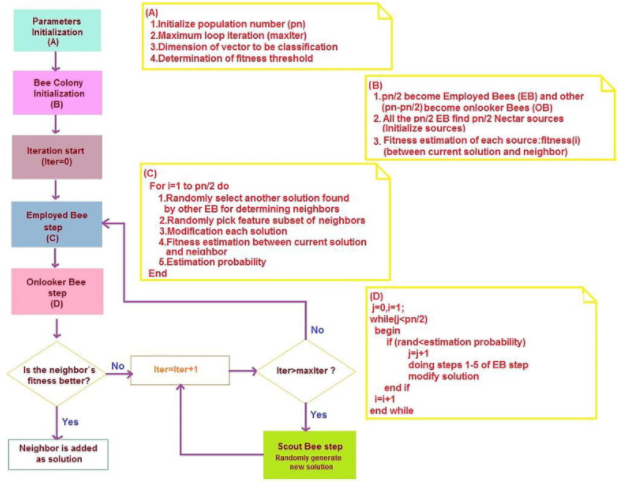
**3. Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms (Mazini, M., Shirazi, B., & Mahdavi, I., 2018).**

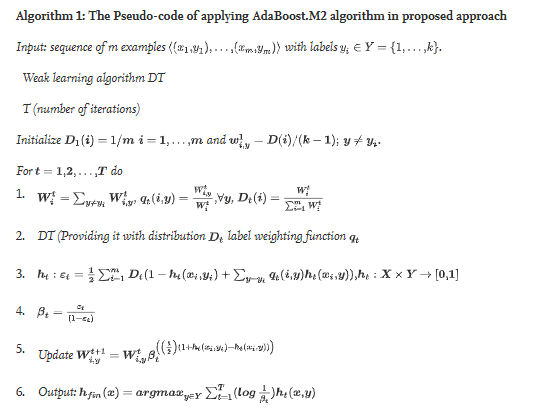
*i) Technique/algorithm used and why it was chosen (motivation)*

This paper proposes the use of artificial bee colony (ABC) for feature selection and AdaBoost for evaluation and classification of features in order to gain a high detection rate with low false positives.

The main aim of this study is to reduce the number of false alarm reports of intrusion to the network that IDSs report due to the high volume of network data, i.e. to propose a model that can handle large amounts of network data without unreasonable increase in the number of false positives.

*ii) Architecture/ model/pseudocode developed*



**

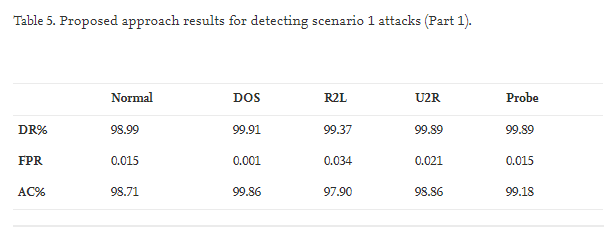
*iii) Datasets analyzed in the paper with the performance results*

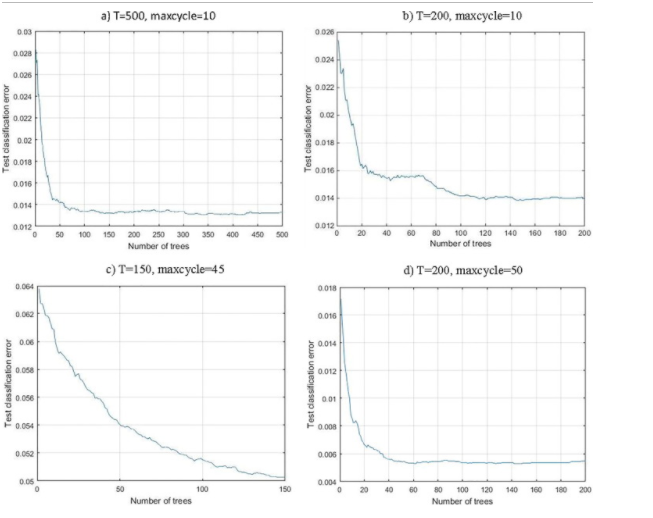
1. NSL-KDD

* Canadian Institute for Cybersecurity in 2009
* Improvement on the KDD CUP 99
* Attacks - DoS, privilege escalation (remote-to-local and user-to-root), probing
* <https://www.unb.ca/cic/datasets/nsl.html>

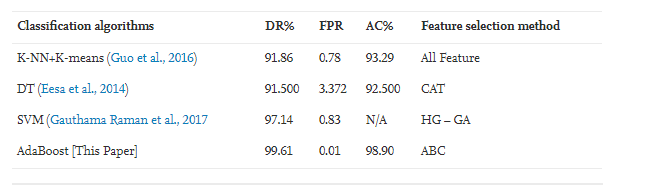
1. ISCXIDS2012

* Canadian Institute of Cybersecurity in 2012
* Captured Friday, 11/6/2010 to Thursday 17/6/2010
* consists of labeled network traces, including full packet payloads in pcap format, which along with the relevant profiles are publicly available for researchers
* Attacks -Infiltrating the network from the inside; HTTP DoS; DDoS using an IRC botnet; SSH brute force
* <https://www.unb.ca/cic/datasets/ids.html>



**

*iv) Any comparison done with the previous techniques to specify that the proposed method is superior*



This table shows that the proposed algorithm in this paper performed much better than the other studied algorithms on the same dataset. Improvement in detection rate and decrease in false positive rate is considerable between the proposed algorithm and the other 3 methods.

**4. Ada-Boosted Locally Enhanced Probabilistic Neural Network for IoT Intrusion Detection. (Jan, T., 2018)**

*i) Technique/algorithm used and why it was chosen (motivation)*

The technique used here is AdaBoosting coupled with an intelligent and compact Probabilistic Neural Network which integrates locally enhanced semiparametric base classifiers. The paper focuses on intrusion detection for systems in IoT environments.

The main motivation of the proposed model is to significantly reduce the computational complexity as machine learning systems generally require high computational complexity for higher rates of accuracy which is infeasible due to the limited computational power and real-time response requirements of IoT systems.

*ii) Architecture/ model/pseudocode developed*

The proposed model is a locally fine-tined semiparametric learning model in which the decision space is locally segmented through the process of semi-supervised learning, where each feature segment is assigned with a fine-tuned simple parametric base classifier. An approximation of a much larger non-parametric neural network model will be done by combining the locally fine tuned parametric models.

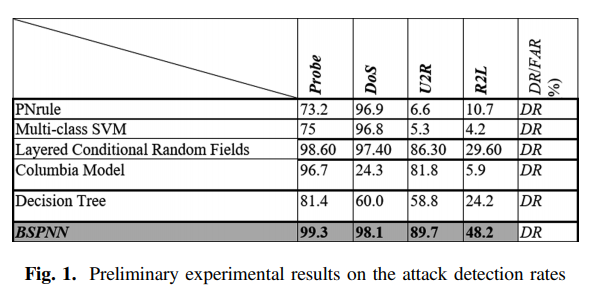
The Adaptive Booster iteratively produces base hypotheses on a weighted training dataset. The weights are updated adaptively based on the classification performance of component hypotheses. The generated hypotheses are then integrated via a weighted sum based on their diversity. The Modified Probabilistic Classifier serves as the base learner which can be trained repeatedly by the AdaBooster to obtain the hypothesis. In each boosting iteration, a base hypothesis is created with associated accuracy and diversity measures. From this information, the data weights are updated for the next iteration and the final weighting of that hypothesis in the joint classification is computed.

*iii) Datasets analyzed in the paper with the performance results*

KDD99 dataset -

* Each record : 41 features, one labelled data
* Attacks - 40 types in 4 categories : PROBE, DoS, User to Root, Remote to Local

*iv) Any comparison done with the previous techniques to specify that the proposed method is superior*



The result shows that the BSPNN model had a significant improvement over other models that were included in the paper.

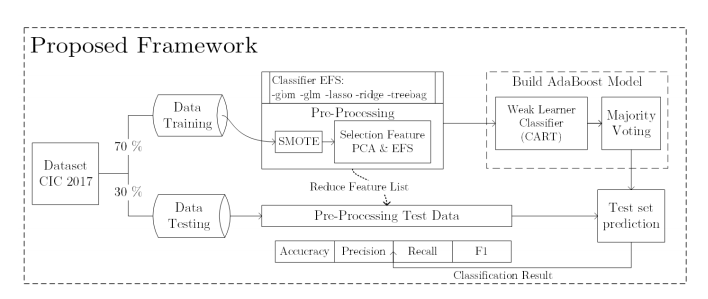
**5. Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset (Yulianto, A., Sukarno, P., & Suwastika, N. A. ,2019)**

*i) Technique/algorithm used and why it was chosen (motivation)*

This paper considers the use of Synthetic Minority Oversampling Technique (SMOTE), Principal Component Analysis (PCA), and Ensemble Feature Selection (EFS) to improve the performance of AdaBoost-based Intrusion Detection System (IDS)

Due to imbalance of training data and inappropriate selection of classification methods, the performance of AdaBoost has room for improvement. The main motivation for this research is to improve the performance of AdaBoost.

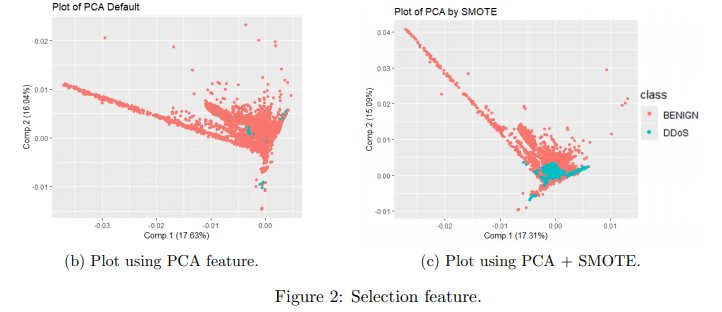
*ii) Architecture/ model/pseudocode developed*



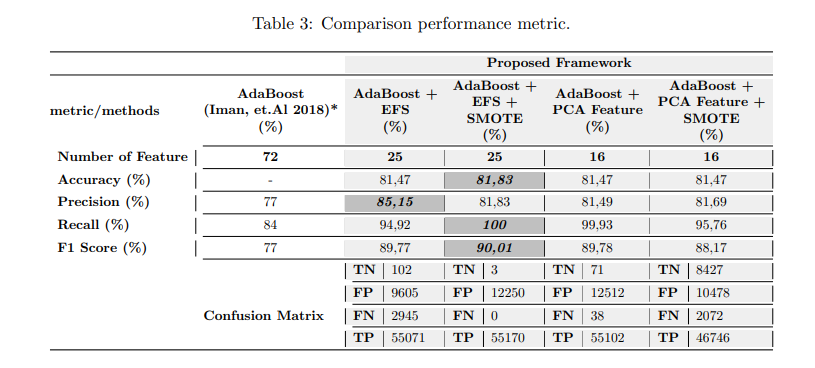
*iii) Datasets analyzed in the paper with the performance results*

CIC IDS 2017 dataset

* Developed by the Faculty of Computer Science, University of New Brunswick in 2017.
* A refinement of the ISCX 2012 datase
* CIC IDS 2017 consist of 5 days of data collection with 225,745 packages with over 80 features and gathered more than seven days of network activity
* Attacks - Brute Force Attack, Heart Bleed Attack, Botnet, DoS Attack, DDoS Attack, Web Attack, and Infiltration Attack. In this paper, analysis of DDoS was conducted



*iv) Any comparison done with the previous techniques to specify that the proposed method is superior*



The given results show that AdaBoost + EFS + SMOTE outperformed the other methods in all areas except for precision, in which it performed second to AdaBoost + EFS

**Comparative Study of results from the papers**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl. No** | **Paper, Author(s), Year** | **Algorithm or Model used** | **Comparison with other Algorithms** | |
| 1. | Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost.  (Shahraki, A., Abbasi, M., Haugen, Ø., 2020) | Real AdaBoost, Gentle AdaBoost and Modest AdaBoost | **Algorithm**  Real AdaBoost  Gentle AdaBoost  Modest AdaBoost  Decision Tree  Ramp Loss K-SVCR  Naive Bayes Probing | **Detection Rate**  ***99.02%***  ***99.02%***  98.46%  98.40%  98.86%  98.42% |
| 2. | Semi-Supervised tri-Adaboost algorithm for network intrusion detection (Yuan, Y., Huo, L., Yuan, Y., Wang, Z., 2019) | Semi-  Supervised tri-adaboost algorithm | **Algorithm**  Semi-Supervised tri-adaboost algorithm  Semi-supervised SVM | **Result**  DR - 89.05%  FPR - ***7.54%***  PR - ***98.54%***  DT ***- 7.34s***  DR - 97.34%  FPR - 8.65%  PR - 92.35%  DT - 98.14s |
| 3. | Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms (Mazini, M., Shirazi, B., & Mahdavi, I., 2018). | Artificial Bee Colony feature selection and Adaboost classification | **Algorithm**  K-NN+K-means with all features  Decision Tree with CAT feature selection  Support Vector Machine with HG-GA feature selection  Adaboost with ABC feature selection | **Result**  DR% - 91.86  FPR - 0.78  AC% - 93.29  DR% - 91.50  FPR - 3.372  AC% - 92.50  DR% - 97.14  FPR - 0.83  AC% - N/A  DR% - ***99.61***  FPR - ***0.01***  AC% - ***98.90*** |
| 4. | Ada-Boosted Locally Enhanced Probabilistic Neural Network for IoT Intrusion Detection | Compact PNN with adaboost classification (BSPNN) | **Algorithm**  PNrule  Multi-class SVM  Layered Conditional Random Fields  Columbia Model  Decision Tree  BSPNN | **Result (DR%)**  PROBE - 73.2  DoS - 96.9  U2R - 6.6  R2L - 10.7  PROBE - 75  DoS - 96.8  U2R - 5.3  R2L - 4.2  PROBE -98.60  DoS - 97.40  U2R - 86.30  R2L - 29.60  PROBE - 96.7  DoS - 24.3  U2R - 81.8  R2L - 5.9  PROBE -81.4  DoS - 60.0  U2R - 58.8  R2L - 24.2  PROBE - ***99.3***  DoS - ***98.1***  U2R - **89.7**  R2L - ***48.2*** |
| 5. | Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset (Yulianto, A., Sukarno, P., & Suwastika, N. A. , 2019) | Use of combinations of Synthetic minority oversampling technique (SMOTE), Principal Component Analysis (PCA), Ensemble Feature Selection (EFS) with AdaBoost Classification | **Algorithm**  AdaBoost  AdaBoost + EFS  AdaBoost + EFS + SMOTE  AdaBoost +PCA  AdaBoost + PCA + SMOTE | **Result (%)**  Accuracy -N/A  Precision - 77  Recall - 84  F1 Score - 77  Accuracy - 81.47  Precision - ***85.15***  Recall - 94.92  F1 Score -90.01  Accuracy - ***81.83***  Precision -  Recall - ***100***  F1 Score - ***90.01***  Accuracy - 81.47  Precision - 81.49  Recall - 99.93  F1 Score -89.78  Accuracy - 81.47  Precision - 81.69  Recall - 95.76  F1 Score -88.17 |

**References**

[1] Shahraki, A., Abbasi, M., & Haugen, Ø. (2020). Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. Engineering Applications of Artificial Intelligence, 94, 103770. doi:10.1016/j.engappai.2020.103770

[2] Moustafa, Nour, and Jill Slay. ["UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)."](https://ieeexplore.ieee.org/abstract/document/7348942) Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

[3] Yuan, Y., Huo, L., Yuan, Y., & Wang, Z. (2019). Semi-supervised tri-Adaboost algorithm for network intrusion detection. International Journal of Distributed Sensor Networks, 15(6), 155014771984605. doi:10.1177/1550147719846052

[4] Mazini, M., Shirazi, B., & Mahdavi, I. (2018). Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms. Journal of King Saud University - Computer and Information Sciences. doi:10.1016/j.jksuci.2018.03.011

[5] Jan, T. (2018). Ada-Boosted Locally Enhanced Probabilistic Neural Network for IoT Intrusion Detection. Complex, Intelligent, and Software Intensive Systems, 583–589. doi:10.1007/978-3-319-93659-8\_52

[6] Yulianto, A., Sukarno, P., & Suwastika, N. A. (2019). Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset. Journal of Physics: Conference Series, 1192, 012018. doi:10.1088/1742-6596/1192/1/012018