Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 5, July, 2021: 3987-3993

Research Article

An Intelligent Method For Network Intrusion Detection System

Noel Jackson,

M.Tech In Network Engineering

Department Of Information Technology,
Rajagiri School Of Engineering & Technology,
Kerala, India

Mail: Noeljack97@Gmail.Com

Ms. Divya James,

Assistant Professor

Department Of Information Technology,

Rajagiri School Of Engineering & Technology,
Kerala, India

Mail:Divyaj@Rajagiritech.Edu.In

Abstract

In Recent Decades, With Rapid Urbanization And Technological Advancements, Millions Of Smart Devices Have Been Flooding The Market. With The Rapidrise In Internet-Based Devices, The Number Of Which Is Anticipated To Reach 50 Billion By The End Of The Decade, Makes Network Security Attacks Essentially A Global Problem. Intrusion Detection Systems Are Systems That Use Techniques To Secure Networks And Devices Against Invasion In Progress Or Other Illegal Activities On A Network. To Date, Existing Ids Designs Have Been Reported With Poor Performance, Including High False Positive Rates And Low Accuracy. Hence, One Of The Most Important Requirements For Designing An Ids Is Improved Detection Performance. For This, A Hybrid Intelligent System Is Proposed That Uses Ga And Fcm In Combination With Machine Learning Algorithms Like Rf, Dt And Gb.

Keywords - Intrusion Detection, Ga, Fcm, Random Forest.

INTRODUCTION

In This Day And Age, Where Technological Advancements Happen At The Blink Of An Eye, We Find Ourselves Surrounded By Internet Connected Devices At All Times. Millions Of These Devices Are Affected By Malware In Some Way Or The Other, Whilst Thousands Of Devices Lie Dormant As Bot Devices For Distributed Attacks. The Rise In Malware Has Been Tremendous In The Last Few Years And It’s Still Growing Exponentially. It Leads To The Loss Of Data And Breach In Privacy.

There Are Many Ways To Cope With This Issue, One Of The Most Effective Way Is To Use Network Intrusion Detection Systems (Nids). Intrusion Detection Systems Are Systems That Use Techniques To Secure Networks And Devices Against Invasion In Progress Or Other Illegal Activities On A Network. Intrusion Detection Systems Can Be Categorized Into Two: Host Based Ids And Network Based Ids. As The Name Suggests, Host-Based Ids Use System Log Files To Analyze Intrusion Detection And Network-Based Ids Uses Network Behaviours. There Are Two Forms Of These Systems, Anomaly Based And Signature-Based Ids, Which Are Based On The Normal And Abnormal Patterns In The Network. A Signature-Based Ids Detects An Intrusion Directly When Predefined Abnormal Network Activities Are Detected, While An Anomaly-Based Ids Analyzes Normal Network Behaviours To Determine Whether An Intrusion Has Occurred Or Not. Existing Ids Designs Have Traditionally Reported Poor Performance, Including Low Accuracy And Very High False Positive Rates. Hence, Improved Detection Performance Is One Of The Determining Requirements In The Designing Of The Ids. To Cope With This Issue, Most Studies Have Focused On Designs That Use Machine Learning Techniques, Which Rely Mainly On Supervised And Unsupervised Methods To Learn Representative Patterns In Network Intrusions. Some Of The Most Common Ml Techniques Are Naïve Bayes, Decision Tree, Support Vector Machines, Random Forest, And Gradient Boosting Algorithm. However, Existing Ml Algorithms Use Only A Small Amount Of Input Data As Utilizing A Very Large Data Set Is A Time-Consuming Process. The Dimensionality And Nonlinear Characteristics Of Large Datasets Usually Make Them Unfit In The Implementation Of Ml Methods To Solve Multiple Classification Tasks, Thereby Reducing The Overall Performance Of The System. Hence, A Feature Selection Utility Can Remove Irrelevant Features From The Feature Set And Improve The Overall Learning Process.

In This Work, We Use Ga And Fcm For The Creation Of Ifs And This Ifs Is Passed Onto Machine Learning Algorithms For Classification Of The Data Into Malignant And Benign, Which Uses Machine Learning.

RELATED WORKS

In [1] A Novel Approach For Network Intrusion Detection Using A Combination Of Genetic Algorithm And Multi-Layered Perceptron Network. The Implementation Is Done On Python 3.6 And Nsl-Kdd Dataset. In This Research Paper, A New Approach For Feature Extraction Is Demonstrated And The Accuracy Of The Outcome Is Evaluated On The Benchmark Nsl-Kdd Data. It Explored Ga And Mlp For Extracting The Features. Results Revealed That The Accuracy Of The System In Detecting Normal, U2r, R2l, Probe And Dos Classed Based On The Extracted Features Is Much Higher Than That Of Previously Proposed Systems And Also It Is Time Efficient Due To The Utilization Of A Reduced Number Of Features. This Approach Provided A High Accuracy On Dos And R2l Classes. This Suggests That If With A Greater Number Of Populations In The Ga And With A Greater Number Of Iterations Accuracy Can Be Achieved In Other Classes As Well. The Proposed Feature Extraction Methodology Is Based On A Combination Of Genetic Algorithms And Multi-Layered Perception Model. The Implementation Segment Is Done On Python3.6 And Nsl-Kdd Dataset Is Used To Test The Accuracy Of Proposed Work Methodology.

In Paper [2] It Proposes A Novel Approach For Intrusion Detection System Based On Sampling With Least Square Support Vector Machine (Ls-Svm). Decision Making Is Performed In Two Stages.

In The First Stage, The Whole Dataset Is Divided Into Some Predetermined Arbitrary Subgroups. The Proposed Algorithm Selects Relevant Samples From These Subgroups Such That The Samples Reflect The Entire Dataset. An Allocation Scheme Based On The Variability Of The Observations Within The Subgroups Is Then Developed. In The Second Stage, Least Square Support Vector Machine (Ls-Svm) Is Applied To The Extracted Samples To Detect Intrusions. The Proposed Algorithm Is Also Known As Optimum Allocation-Based Least Square Support Vector Machine (Oals-Svm) For Ids. To Demonstrate The Effectiveness Of The Proposed Method, The Experiments Are Carried Out On Kdd 99 Database Which Is Considered A De Facto Benchmark For Evaluating The Performance Of Intrusions Detection Algorithm. All Classes Are Tested, And The Proposed Architecture Obtains The Needed Accuracy And Efficiency. Finally, The Usability Of The Proposed Algorithm For Incremental Datasets Is Also Shown.

In Paper [3] Takes Nsl-Kdd Data Set As The Base Dataset, Then It Analyzes The Latest Progress And Existing Works In The Field Of Intrusion Detection Technology, And Finally It Proposes An Adaptive Ensemble Learning Model. By Adjusting The Proportion Of Training Data And Setting Up Multiple Decision Trees, It Constructs A Multitree Algorithm. To Improve The Overall Detection Effect, It Chose Several Base Classifiers, Including Decision Tree, Random Forest, Knn, Dnn, And Design An Ensemble Adaptive Voting Algorithm. It Uses Nsl-Kdd Test+ To Verify The Approach, The Accuracy Of The Multitree Algorithm Is 84.2%, While The Final Accuracy Of The Adaptive Voting Algorithm Reaches 85.2%. Compared With Other Research Papers, It Is Proved That This Ensemble Model Effectively Improves Detection Accuracy. In Addition, Through The Analysis Of Data, It Is Found That The Quality Of Data Features Is An Important Factor To Determine The Detection Effect. The Adaptive Ensemble Learning Model Designed In This Paper Chooses Common Machine Learning Algorithms Such As Decision Tree, Svm (Support Vector Machines), Logical Regression, Knn (K-Nearest Neighbours), Adaboost, Random Forest And Deep Neural Network As Alternative Classifiers. Five Voting Classifiers Are Selected Through Comparative Tests. Then, By Adjusting The Proportion Of Samples, Setting Data Weights, Multi-Layer Detection And Other Combined Method To Boost The Detection Effect Of Each Algorithm. Finally, The Adaptive Voting Algorithm With Different Class-Weights Is Used To Obtain The Optimal Detection Results. [3] Proposes That The Diverse Wireless Network Traffic Attack Characteristics Has Led To Several Problems What Traditional Intrusion Detection Technology With High False Positive Rate, Low Detection Efficiency, And Poor Generalization Ability Can’t Properly Detect. To Enhance The Security And Improve The Detection Ability Of Malicious Intrusion Behaviour In A Wireless Network, This Paper Proposes A Wireless Network Intrusion Detection Method Based On Improved Convolutional Neural Network (Icnn). First, The Network Traffic Data Is Categorized And Pre-Processed, Then Modelled The Network Intrusion Traffic Data By Icnn. The Low-Level Intrusion Traffic Data Is Abstractly Represented As Advanced Features By Cnn, Which Extracted Autonomously The Sample Features, And Optimizing Network Parameters By Stochastic Gradient Descent Algorithm To Converge The Model. Finally, We Conducted A Sample Test To Detect The Intrusion Behaviour Of The Network. The Simulation Results Show That The Method Proposed In Our Paper Has Higher Detection Accuracy And True Positive Rate Together With A Lower False Positive Rate. The Test Results On The Test Set Kddtest + In Our Paper Show That Compared With The Traditional Models, The Detection Accuracy Is 8.82% And 0.51% Higher Than That Of Lenet-5 And Dbn, Respectively, And The Recall Rate Is 4.24% And 1.16% Higher Than That Of Lenet-5 And Rnn, Respectively, While The False Positive Rate Is Lower Than The Other Three Types Of Models. It Also Has A Big Advantage Compared To The Idabcnn And Nidmbcnn Methods. This Paper Considers Using The End-To-End Semi-Supervised Network Training Classifier Of Convolutional Neural Network (Cnn), And The Multi-Layer Feature Of Cnn To Detect Network, Learn The Sample Features And Discover The Rules In The Data Training Process To Simplify The Implementation Process.

In [5], It Proposes A Novel Algorithm For A Network Intrusion Detection System (Nids) Using An Improved Feature Subset Selected Directly By A Genetic Algorithm (Ga)-Based Exhaustive Search And Fuzzy C-Means Clustering (Fcm). The Algorithm Identifies The Bagging (Bg) Classifier And The Convolutional Neural Network (Cnn) Model As An Effective Extractor By Implementing The Ga In Combination With 5-Fold Cross Validation (Cv) To Select The Cnn Model Structure. The Deep Feature Subset Is Extracted By The Selected Cnn Model Which Is Then Put Into The Bg Classifier To Validate The Performance With The 5-Fold Cv. The High-Quality Feature Set So Obtained By The Three-Layered Feature Construction Using The Ga, Fcm, Cnn Extractor, And A Hybrid Cnn And Bg Learning Method Significantly Improves The Final Detection Performance. Also, The Highly Reliable Validation Performance Results Achieved By The 5-Fold Cv Procedure For The Proposed Algorithm Imply A Well-Fitted Application In A Practical Computer Network Environment Nids.

Iii. Proposed System Design

The Proposed Architecture Mainly Focuses On Feature Selection And Model Generation. The Extent Of Detection Accuracy Is Closely Related To The Feature Selection. The Figure Illustrates The System Architecture And Components That Are Involved In The System. There Are Mainly 3 Modules In The Architecture, Module 1 Checks For Relevant Features In The Dataset And Generates And Improved Feature Subset (Ifs). Module 2 Selects The Model Using A Convolutional Neural Network. Then These Two Modules Are Combined To Form

A Single Output. Module 3 Finally Validates The Proposed Architecture And Produces The Accuracy And Overall Effectiveness Of The Model.

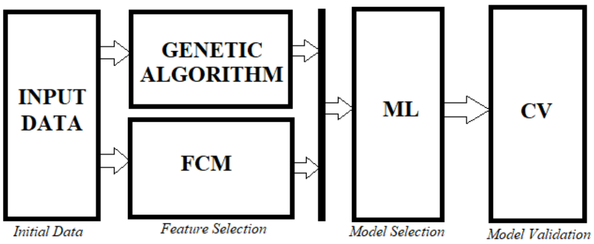


Figure 1: System Architecture

Phases Of The System

In This Section, Detailed Information About The Components Of Our Proposed System Is Given.

A. Data Preprocessing

The Dataset Consists Of Around 494021records Which Consist 4 Network Attacks, Dos, R2l, U2r, And Probe. These Are Divided Into Testing And Validation And Placed In Separate Folders. It Takes In The Dataset To Be Analyzed, Creates A Dictionary For The Attack Types, Then It Finds And Corrects Missing Values And Finally Correlated Data Are Removed From The Dataset. The Cleaning Process Reduces The Feature Count From 43 To 33 After The Process Is Completed.

B. Feature Selection

This Is A 2-Part Process, First The Original Dataset Is Passed Onto A Genetic Algorithm For Computing Relevant Features From The Dataset, Which Approximately Reduces The Feature Count From 33 To 20.

The Genetic Algorithm Has A 5-Step Process, It First Generates An Initial Population Using A Random Function Generator, Then Mating Pool Is Selected From The Dataset Based On A Fitness Function, Support Vector Machine Is Used As The Fitness Function Here. Then Crossover Process Is Performed To Generate Offspring's, These Offspring Are Again Checked Against The Fitness Function To Test The Efficiency Of The System.

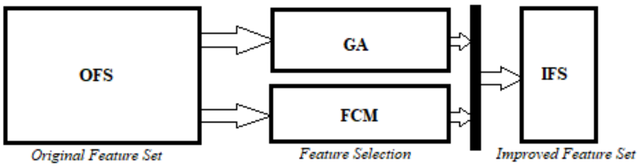


Figure 2: Feature Selection

Mutation Is Performed At Random Intervals To Reduce Premature Convergence Of The Evolution Process. Finally, The Best Outputs Are Selected From The Population Which Form The Feature Subset For Further Evaluation Of The Dataset. The Next Part In Module 2 Is Passing The Dataset To A Fuzzy C Means Clustering Algorithm To Reduce Variance And To Compute Additional Features That May Have Been Excluded By The Genetic Algorithm. For Fcm, First Weights Are Initialized Using A Random Function Generator, Then Centroids Are Computed For Each Element In The Dataset. Now The Weights Are Updated Based On The Distance Between Elements In The Dataset. Finally, The Elements Are Clustered Into 2 Groups. The Smaller Cluster Is Selected To Reduce The Overall Computational Cost. The Outputs Of The Fcm And Ga Are Compared And Combined To Give The Improved Feature Subset (Ifs). The Ifs Is Passed To Machine Learning Algorithms Like Random Forest (Rf), Naïve Bayes (Nb) And Gradient Boosting (Gb) Classifiers. The Best Machine Learning Algorithm Is Selected As The Classification Model. This Is The Feature Selection Module.

C. Rf

Random Forests Or Random Decision Forests Are An Ensemble Learning Method For Classification, Regression And Other Tasks That Operates By Constructing A Multitude Of Decision Trees At Training Time.

For Classification Tasks, The Output Of The Random Forest Is The Class Selected By Most Trees. For Regression Tasks, The Mean Or Average Prediction Of The Individual Trees Is Returned. Random Decision Forests Correct For Decision Trees' Habit Of Overfitting To Their Training Set. Random Forests Generally Outperform Decision Trees, But Their Accuracy Is Lower Than Gradient Boosted Trees. However, Data Characteristics Can Affect Their Performance. Random Forests Are Frequently Used As "Blackbox" Models In Businesses, As They Generate
Reasonable Predictions Across A Wide Range Of Data While Requiring Little Configuration.

D. Naïve Bayes

Naive Bayes Is A Simple Technique For Constructing Classifiers Models That Assign Class Labels To Problem Instances, Represented As Vectors Of Feature Values, Where The Class Labels Are Drawn From Some Finite Set. There Aren’t Many Algorithms To Train Such A Classifier, It Is All Based On A Common Principle: All Naive Bayes Classifiers Assume That The Value Of A Particular Feature Is Independent Of The Value Of Any Other Feature, Given The Class Variable. For Example, A Chair May Bea Stool If It Is Small, Round, And About 50 Cm In Height. A Naive Bayes Classifier Considers Each Of These Features To Contribute Independently To The Probability That This Chair Is A Stool Or A Couch, Regardless Of Any Possible Correlations Between The Colour, Roundness, And Height Features.

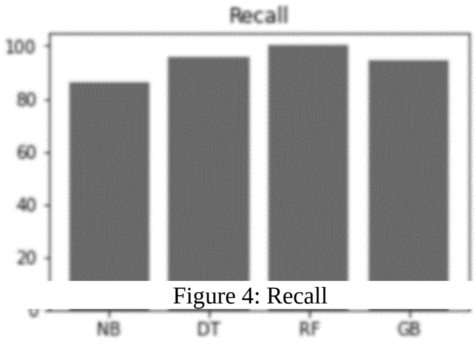


Figure 4: Recall

For A Lot Of Probability Models, Naive Bayes Classifiers Can Be Trained Very Efficiently In A Supervised Learning Setting. In Many Practical Applications, Parameter Estimation For Naive Bayes Models Uses The Method Of Maximum Likelihood; In Other Words, One Can Work With The Naive Bayes Model Without Accepting Bayesian Probability Or Using Any Bayesian Methods. Despite Their Naive Design And Apparently Oversimplified Assumptions, Naive Bayes Classifiers Have Worked Quite Well In Many Complex Real-World Situations. In 2004, An Analysis Of The Bayesian Classification Problem Showed That There Are Sound Theoretical Reasons For The Apparently Implausible Efficacy Of Naive Bayes Classifiers. Still, A Comprehensive Comparison With Other Classification Algorithms In 2006 Showed That Bayes Classification Is Outperformed By Other Approaches, Such As Boosted Trees Or Random Forests. An Advantage Of Naive Bayes Is That It Only Requires A Small Number Of Training Data To Estimate The Parameters Necessary For Classification.

E. Model Validation

In Machine Learning, Model Validation Is Often Considered As The Process Where A Trained Model Is Evaluated On A Testing Dataset. The Testing Dataset Is Usually A Separate Portion Of The Same Dataset From Which The Training Dataset Was Derived. The Objective Behind This Rationale Is Testing The Generalization Capability Of The Trained Model.

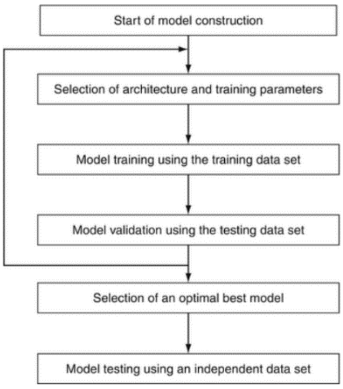

Start of model construction
Selection of architecture and training parameters
Model training using the training data set
Model validation using the testing data set
Selection of an optimal best model
Model testing using an independent data set

Figure 3: Model Validation

RESULTS

This Study Demonstrates An Effective Algorithm Which Is Practically Possible To Implementon A Practical Ids And It Plays A Crucial Role In Detecting Potential Illegal Activities Over Computer Networks. Although Countless Other Methods Have Been Employed For Multiple Intrusion Problem Classification Tasks, Their Overall Performance Was Measured To Be Relatively Poor, Which Makes These Ids Designs Difficult To Apply For Practical Network Environments. The Ml Technique, Which Has Been Considered For Ids Design, Can Be Viewed As Having Two Parts: Feature Extraction And Classification.

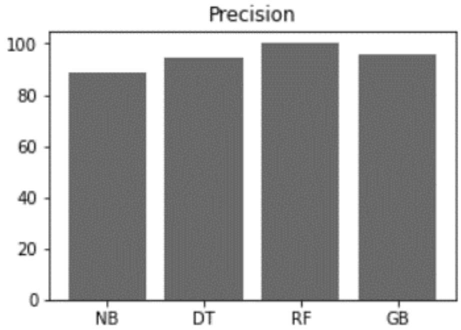
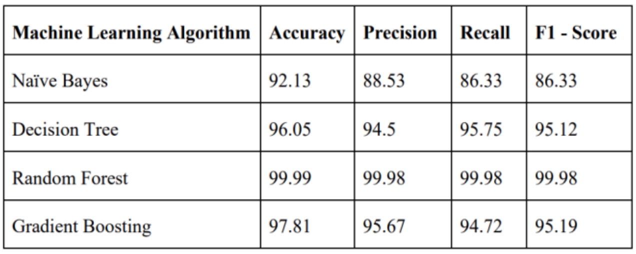


Figure 5: Precision

The First Part Plays The Role Of Feature Extraction Which Is Done Using A Combination Of Ga And Fcm And The Second Part Plays The Role Of Classification Of Network Data To Malignant And Benign Categories. The Disadvantage Of Using Just Fcm Or Ga Is That The Features Extracted May Not Always Be Relevant In All Scenarios. Thus, A Two-Layered Feature Construction Strategy Is Introduced, Including The Feature Selection To Select The Relevant Features From The Ofs For The Ml Input.



Machine Learning Algorithm	Accuracy	Precision	Recall	F1 - Score
Naïve Bayes	92.13	88.53	86.33	86.33
Decision Tree	96.05	94.5	95.75	95.12
Random Forest	99.99	99.98	99.98	99.98
Gradient Boosting	97.81	95.67	94.72	95.19

Figure 6: Performance Comparison

Compared To The Previously Mentioned Studies, This Method Proposes A Two-Layered Feature Construction To Increase The Feature Quality, Which Is Definitely Effective In Improving The Final

Classification Performance. Indeed, The Ga-Based Feature Selection Plays A Role Of The First Layer To Select The Most Informative Feature Subset, Which Is Then Improved By Adding More Efficient Features Computed By The Second Layer Of Feature Improvement Using Fcm. The Validation Performance Of The Ml Classifiers Is By Using The Ifs, Which Is Considered As The Two-Layered Feature Construction. Therefore, The Rf Model Is An Effective Ids Application As The Extractor Ml Classifier. Furthermore, The Proposed Algorithm Using The Ifs Produces Better Performance Than That Using The Ofs. Obviously, The Ml Classifiers' Validation Performance Is Relatively Low Compared To The Proposed Algorithm. These Results Imply That The Ga And Fcm Are Effective In Constructing An Informative Ifs.

CONCLUSIONS

Correct Ids Classification Of Dangerous Computer Network Attacks Is Important For Rapid Responses To These Attacks In Terms Of Network Security. An Efficient Algorithm For Deployment In A Practical Nids Can Be Implemented, Using A Combination Of An Ifs And An Ml Extractor. The Ifs Can Be Carefully Constructed By The Ga To Select The Most Informative Feature Set From The Ofs And The Fcm To Compute Additional Features. Moreover, The Ifs Can Then Be Used As The Input To The Rf Model, Which Selects Various Features Structures Generated By The Ga And Fcm. The Simulation Is Expected To Show The Successful Utility Of Hybrid Algorithm For Cyber Security. Indeed, This Would Be An Effective Method To Make Massive Intrusion Data More Separable By The Use Of Ml Techniques. As A Result, It Is Expected That Deployment Of The Proposed Algorithm Over The Practical Internet Systems Would Improve The Computer Network Security Against The Illegal Activities.

REFERENCES

1. Joydev Ghosh, Divya Kumar, Rajesh Tripathi: "Features Extraction For Network Intrusion Detection Using Genetic Algorithm (Ga)", Studies In Computational Intelligence, Springer (2020).
2. Kabir, J. Hu, H. Wang, G. Zhuo: "A Novel Statistical Technique For Intrusion Detection Systems", Future Gener. Comput. Syst. 79 (2018) 303–318.
3. X. Gao, C. Shan, C. Hu, Z. Niu, Z. Liu: "An Adaptive Ensemble Machine Learning Model For Intrusion Detection", Ieee Access (2019) 82512–82521.
4. Yang, F. Wang: "Wireless Network Intrusion Detection Based On Improved Convolutional Neural Network", Ieee Access 7 (2019) 64366–64374.
5. A.S. Sohal, R. Sandhu, S.K. Sood, V. Chang, A Cybersecurity Framework To Identify Malicious Edge Device In Fog Computing And Cloud-Of-Things Environments, Comput. Secur. 74 (2018) 340–354. S. Revathi, A. Malathi, A Detailed Analysis On Nsl-Kdd Dataset Using Various Machine Learning Techniques For Intrusion Detection, Int. J. Eng. Res. Technol. 2 (2013) 1848–1853.
6. T.A. Tchakoucht, M. Ezziyyani, Multilayered Echo-State Machine: A Novel Architecture For Efficient Intrusion Detection, Ieee Access 6 (2018) 72458–72468.
7. R. Yahaloma, A. Sterena, Y. Nameria, M. Roytmana, A. Porgadorb, Y. Elovici, Improving The Effectiveness Of Intrusion Detection Systems For Hierarchical Data, Knowl. Based Syst. 168 (2019) 59–69.
8. M. Ahsana, M. Mashuria, M.H. Leeb, H. Kuswantoa, D.D. Prastyo, Robust Adaptive Multivariate Hotelling's T2 Control Chart Based On Kernel Density Estimation For Intrusion Detection System, Expert Syst. Appl. 145 (2020).
9. J. Lee, J. Kim, I. Kim, K. Han, Cyber Threat Detection Based On Artificial Neural Networks Using Event Profiles, Ieee Access 7 (2019) 165607–165626.
10. T.Saranyaa, S.Sridevi B, Performance Analysis Of Machine Learning Algorithms In Intrusion Detection System, Procedia Computer Science 127 (2018).
11. Jiadong Ren, Jiawei Guo, Wang Qian, Building An Effective Intrusion Detection System By Using Hybrid Data Optimization Based On Machine Learning Algorithms, Security And Communication Networks Volume 2019, Article Id 7130868.