RESEARCH ARTICLE OPEN ACCESS

A Novel Approach to Prevent Intrusion Detection through Analysis of Machine Learning and Genetic Algorithm

Praveena.S [1], Mrs.V.Vanitha [2]
Department of Computer Science
Krishnaswamy College of Science, Arts & Management for Women, Cuddalore
Tamil Nadu - India

ABSTRACT
Communication systems are vital for effective transfer of information over a network. However, there is many threats in maintaining confidentiality and promoting high-level security of information to increase the quality of network communication and increase the trust of the people involved in it. The Research Thesis is an attempt to implement a hybridized genetic algorithm for enhancing authentication and filtering the traffic so that congestion prevented and complexity reduced. The Algorithm used a renounced dataset that shows transmission of huge amount of data from KDD99. To make the results of the algorithm more effective a new dataset created. The Research focuses on the prediction of Intrusion in Network congestion and exaggerates the presence of intrusion in a network with the use of research tools that can predict the presence of intrusion in a host or network based architecture using Analytics from three proposed algorithms namely Naïve Bayesian Algorithm, Stacking Algorithm forming a hybrid with Genetic Algorithm . The calculations made based on the statistical analysis of data in the network dataset and results predicted using genetic algorithm in Java based environment with Eclipse. The Analytics worked out in Weka on all the algorithms with the variations of outputs in predictions of Intrusions over Networks.
Keywords :— Congestion Prevention, Intrusion Detection, Machine Learning, Genetic Algorithm, refined KDD99 dataset

I. INTRODUCTION

In a World of Communication, Intrusion is always a threat for Global users of the system. Their system might be attacked by intruders through suspicious activity like anomalies or misuse in detection. In the anomalies mode, data auditing will help to detect intruders whereas in misusing mode, signature based Detection is used to differentiate the users of the system [1]. The Second model is much equal to antivirus programs installed in the system. The Misuse of the system can analyse and compare signature based activities with network behaviour. There is a need to identify and differentiate the normal traffic with intrusion in the system. There are other techniques like artificial intelligence that can bring best solution for this problem. However it is highly cost effective. The Rules framed for defining the operations in the Intrusion system can be represented by defining patterns or rulesets that can be derived from network traffic area, logs from systems either host or system dependent.

Intrusion Detection System (IDS) needs sensors to get installed and get executed. TCP/IP headers and the number of bytes act as network sensor monitoring the incoming and outgoing packets during transmission [2]. The host sensor monitors various packets related to memory logs and information related to transformation of data.

The System is developed in a new way where direct communication with the server is not entertained thereby preventing the heavy traffic of network because of just monitoring the systems. Hence it is maintained as a separate infrastructure.

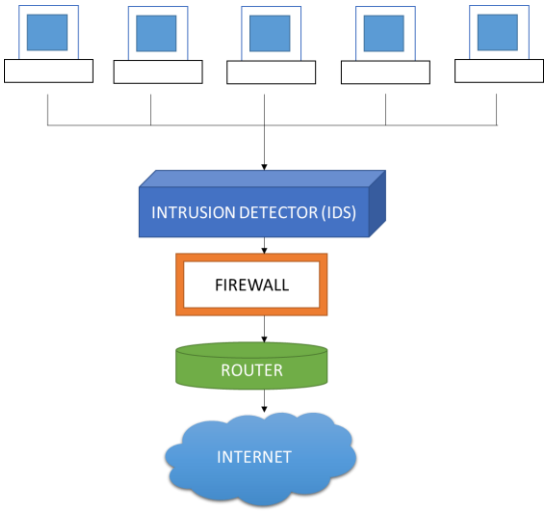

Figure 1 General Intrusion Detection System Architecture

As indicated in the diagram, the intrusion detector acts as a separate infrastructure in no relation with other components of the network systems, firewall, router or internet systems [3]. Intrusion detection of a network based environment infrastructure includes the following features:

1. It helps in analysing & scrutinising user as well as system based activities
2. It can also find system configuration defects or any vulnerabilities in the system

3. It can also understand typical attack patterns in the network based system

4. It can also track user policy and copyright violations in the network

5. It also counts and reports unwanted activities in the system

In general, Intrusion Detection System is based on two major steps. They are

In Host based system, there is passive intrusion detection to enable users to set right passwords [4], inform them of unwanted activities or also advise them on the recommended settings for the system to prevention of intrusion like installing Antivirus or enabling the firewall settings respectively.

In Network based system, the active intrusions are considered. This method enables the system to record every activity of the user and also alarms the unexpected or unwanted activity from unknown users of the system [5]. This is considered as most important as it will be the IDS that will give direct information to the user on active threats of the system which is considered to give harmful effects to the network systems as a whole.

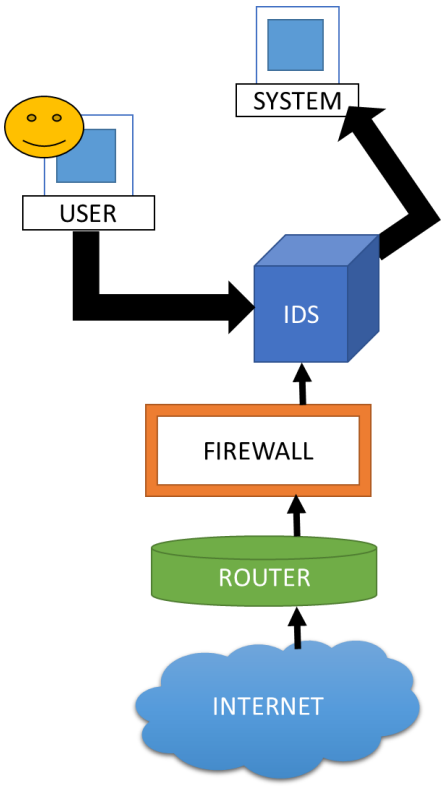

Figure 2 – Intrusion detection Process

The proposed research is based on the Network based infrastructure that is proposed to be implemented using machine learning and genetic algorithms. The research work is explored from the review of literature.

II. REVIEW OF LITERATURE

An Intrusion detection system is one of the important thrust areas in research field for computer science and has high impact in the research field.

1. Networks employs complicated datasets that will challenge the researcher to learn new sequences and get many rapid outputs

2. It is very interesting to evaluate the results of intrusion and predict the results for the same.

Various authors have previously presented their ideas on intrusion detection architecture based on various technologies and research designs.

1. Harendra Vaidya et.al analysed to build two levels of IDS in order to detect abnormal behavior of network traffic and covering fast networks for Intrusion prevention. The work is carried out with Naïve Bayesian algorithm with results sufficient to predict the occurrence of intrusion in computers.
2. Rana Aamir Raza Ashfaq et.al carried out a research with semi-supervised learning algorithm. He utilised unlabelled samples that was supported by supervised learning algorithm to enhance performance of the system.
3. Anthony Melaragno et.al proposed a new model with Rail Radio Intrusion Detection systems (RRIDS) that can identify and monitor various cyber-crimes like command replay, guessing with message modifications. It is variable and can be portable to other systems.
4. Keke Gai et.al presented a reduced version of all major techniques implemented in IDS with new challenging techniques like cloud computing interconnected systems. This framework is considered as a higher level secure component for futuristic technologies with wireless sensors.
5. Abdulla Amin Aburomman et.al took five random subsets from the well-known KDD99 dataset. Ensemble classifiers are created using the new approaches as well as the weighted majority algorithm (WMA) approach. Our experimental results suggest that the new approach can generate ensembles that outperform WMA in terms of classification accuracy.

III. REFINING DATASET WITH ALGORITHMS

Intrusion detection holds a traditional dataset from KDD that contain a total number of 24 training set that also contains an extra 14 attributes in testing dataset [6]. The training dataset is required to be refined using a data mining tool like Rapid Miner that can normalize the data and give effective output dataset that will be the testing dataset in the research.

The detailed description with name is explained with the basic features as underlined below:

A. Basic Features: This method combines all the required properties that could be possible from a normal TCP/IP connection system that can bring implicit delay in the system [7].

B. Feature based on Network Traffic: This feature passively monitors features that can be monitored and computed based on an interval between them [8].

TABLE 1
DERIVED FEATURES OF DATASET

FEATURE NAME	TYPE
duration	Continuous
protocol_type	Discrete
service	Discrete
src_bytes	Continuous
dst_bytes	Continuous
flag	discrete
land	Discrete
wrong_fragment	Continuous
urgent	Continuous

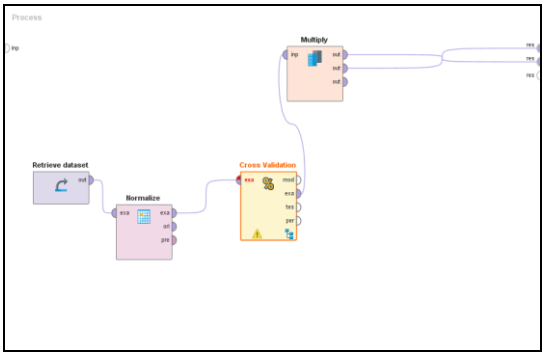

Figure 3: Convert and refine dataset based on KDD

In general, KDD dataset in a refined form has the following advantages:

1. The dataset does not contain repeated records and hence the classifier will not be oriented towards training set based on frequency.
2. Since duplicate records are avoided in the refined set, the research output will be highly acceptable.
3. The shortlisted records in the testing set can present accurate information related to the KDD dataset.
4. Since the testing and training sets are at an acceptable numbers, the research work could be easily completed in quick time and can also present best result that could be more accurate compared to the unrefined dataset.

IV. IMPLEMENTATION AND EVALUATION

After Careful analysis and Review of Literature, the existing system the proposed system was formulated based on the efficiency of the algorithm. The following Techniques were utilized for the effective output of the Intrusion detection system.

The design of algorithm has the following stages with the WEKA Explorer:

1. Dataset loading into system
2. CFS Method to select the right features
3. Predicting the results before selection
4. Predicting the results after selection

The dataset KDD'99 is designated for research as refined dataset, converted into CSV data and loaded into the WEKA tool for evaluation [9].

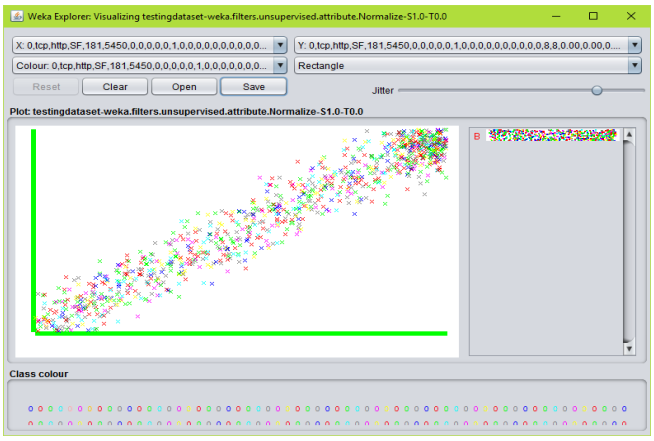

Figure 4 Loading Dataset into WEKA

After loading the dataset into WEKA, the feature selection process using CFS is carried out.

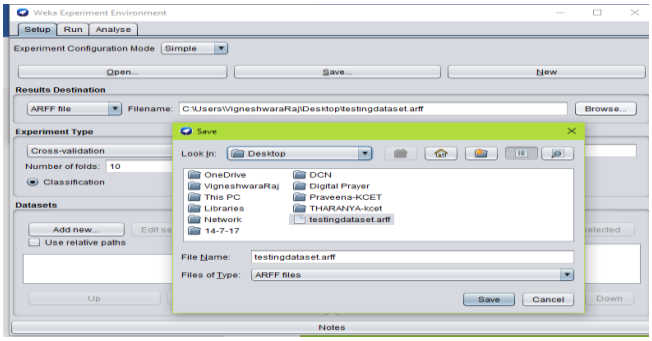

Figure 5 Feature Selection Using CFS Method

After feature selection, the resultant data is viewed in the graphical and tabular form as indicated in the diagram.

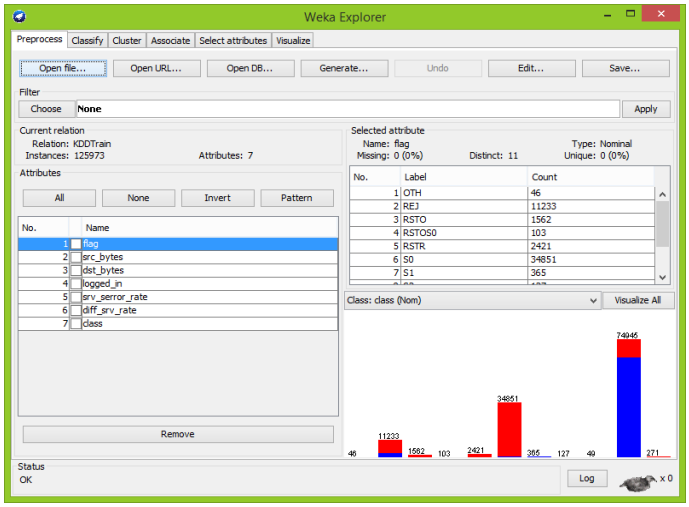

Figure 6 after Feature Selection

Thus the dataset is evaluated in the Explorer by viewing the selected features of the KDD Set. Now the algorithms has to be applied to the dataset to test the intrusions.

4.1 NAÏVE BAYESIAN ALGORITHM

Naïve Bayesian is a classifying algorithm that is based on Mathematical Bayes Theorem where the prediction factors are considered independent of each other [10]. This algorithm considers presence of specific feature is more important without any requirement. Hence it is used along with Intrusion detection systems to enhance prediction.

First, Naïve Bayesian Algorithm is applied to the dataset and the following results are identified.

```
Time taken to build model: 0.28 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      103161               81.8914 %
Incorrectly Classified Instances     22812               18.1086 %
Kappa statistic                          0.6279
Mean absolute error                      0.1809
Root mean squared error                  0.4233
Relative absolute error                 36.3493 %
Root relative squared error             84.8609 %
Total Number of Instances           125973

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.978     0.364      0.755    0.978      0.852      0.937    normal
                 0.636     0.022      0.962    0.636      0.766      0.936    anomaly
Weighted Avg.    0.819     0.205      0.851    0.819      0.812      0.937

=== Confusion Matrix ===

     a     b   <-- classified as
 65865  1478 |     a = normal
 21334 37296 |     b = anomaly
```

Figure 7 NavieBayes-After Feature Selection

Thus it is found that the True Positive Rate is higher comparing to the False Positive rate for intrusion detection in network based systems.

4.2 STACKING CLASSIFIER ALGORITHM

Intrusion detection is based on multiple processes at the same time due to trafficking of data and hence Stacking algorithm plays a significant role to combine multiple classification models through a meta-classifier model [11].

By applying the stacking method, the following results are obtained in WEKA Explorer.

```
Time taken to build model: 0.8 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       67343               53.4583 %
Incorrectly Classified Instances     58630               46.5417 %
Kappa statistic                          0
Mean absolute error                      0.4976
Root mean squared error                  0.4988
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances           125973

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.535     1         0.697      0.5      normal
                 0         0          0         0         0          0.5      anomaly
Weighted Avg.    0.535     0.535      0.286     0.535     0.372      0.5

=== Confusion Matrix ===

     a     b   <-- classified as
 67343     0 |     a = normal
 58630     0 |     b = anomaly
```

Figure 8 Stacking-After Feature Selection

In the above method, it is found that True Positive is equal to the True Negative and hence stands at 50% chances of predicting the right intrusion and false intrusion.

4.3 GENETIC ALGORITHM

Human Genome is a striking factor to identify the possibility of intrusion in normal to distributed systems. In the above model, detection can be made for a small population of distributed systems [12]. In Genetic algorithm, search-based optimisation is rendered and the algorithm helps in best feature selection with hybridisation from other two algorithms.

The dataset is loaded with genetic parameters and tools to identify the prediction accuracy of the intrusion systems.

```
Learned category ----------> anomaly
Selected positive terms size:  5

============= List of the best 10 terms of the selected vocabulary =============
Terms                                Term Score (Chi Square)
"logged_in"                    0.10220428102634625
"src_bytes"                    859.3098320105095
"flag"                  0.0
"diff_srv_rate"                316.61818927967323
"srv_serror_rate"                    631.2418611698608
 ----------------- CLASSIFIER  -----------------

anomaly <--- ("flag")


---------- Learning Results for Category anomaly ----------

#Pos Terms: 1 (# of rules)
#Neg Terms: 0 (negative terms per rule)
|----------------------------------------|----------------------------------------|
|                Train                   |                Test                    |
|----------------------------------------|----------------------------------------|
|    Precision     Recall     FMeasure    |    Precision     Recall     FMeasure    |
|----------------------------------------|----------------------------------------|
       90.73        100.0        95.14    |       0.0  0.0        0.0      |

============== Confusion Matrix ==============

       anomaly normal
       0       47201
       0       66175
```

Figure 9 Genetic Algorithm-After Feature Selection

The result clearly indicates the selection parameters as well as the precision level as normal to indicate accurate presence of intrusion in the dataset.

Thus the dataset is highly refined and gets more accurate results comparing to the former datasets.

After experiments, the results are identified and has to be discussed. Before performing the discussion, few testing methods are used for evaluation of the experiments. The next experiment is conducted with Eclipse software for result analysis and to predict whether there is intrusion or not.

V. RESULT DISCUSSION WITH PARAMETERS

After careful experimentation, various new results are identified and presented as input to the system. The results were identified from WEKA as well as Eclipse for analysis of reports regarding prediction of intrusion in systems. The following results were identified for all the three algorithms based on the following parameters:

1. Accuracy
2. Precision
3. Recall
4. F-Measures
5. FP-Rate
6. Kappa

TABLE 2

The Result analysis of the Existing System

Before---	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**FP-Rate**	**Kappa**
Genetic Algorithm	91.09	98.7	96.15	91.03	8.31	82
NavieBayes	90.38	90.5	90.4	90.42	10.1	80.6
Stacking	53.45	28.6	53.5	37.2	53.5	50.03

The Table indicates high accuracy with genetic algorithm with high values in other parameters as well. The Graphical analysis is also carried out to check whether the genetic algorithm is capable of producing accurate results comparing to other machine learning classifiers.

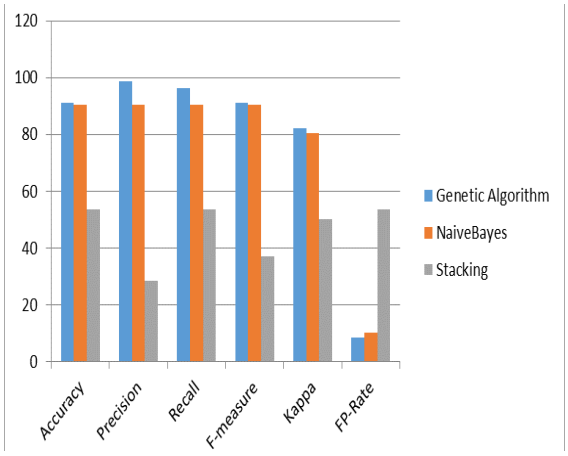

Figure 10 - Graphical analysis of the parameters for prediction analysis

TABLE 3

The Result analysis of the Proposed System

After	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**FP-Rate**	**Kappa**
Genetic Algorithm	97.77	98.38	98.43	98.41	1.1296	94.7
Navie Bayes	81.89	85.1	81.19	81.2	20.5	62.7
Stacking	53	28.6	53.5	37.2	53	50

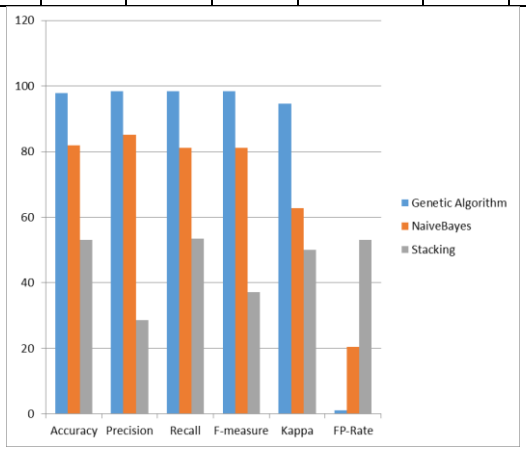

Figure 11 - Graphical analysis of the proposed system

From the graph, it is learnt that the accuracy and precision of the existing system is normal and needs improvement. However, the proposed system showed tremendous improvement for the same as it has the improvised genetic algorithm in its side. The graphs also show that the rate of accuracy and precision along with other factors like Recall, F-Measure, Kappa and FP Rate increases with the proposed system and it lowers down in the existing system for all the three algorithms Naïve Bayesian Algorithm, Stacking algorithm and Genetic Algorithm respectively. The outputs of various Algorithms identified in WEKA tool is represented below:

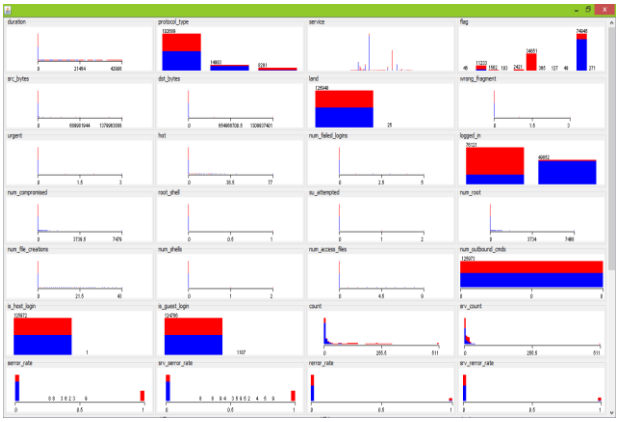

Figure 12 - Output of various algorithms result in graphical format

VI. CONCLUSIONS

The Research work is an extension of various other research works in various stages on Intrusion detection systems to carry over effective protection of network systems from unexpected intrusions from various sources. The prologue and various concepts related to Intrusion detection systems has been discussed along with the motivation for the research work and the contribution of research for the field of computer science. The Result identified as a hybridized projection of Machine Learning and Genetic Algorithm is a new beginning for more and high secure systems in the future.

REFERENCES

[1] Vaidya, H., Mirza, S., & Mali, N. (2016). Intrusion Detection System. International Journal of Advanec Research in Engineering, Science & Technology, 3.

[2] Kukreja, K., Karamchandani, Y., Khandelwal, N., & Jewani, K. (2015). Intrusion Detection System. *International Journal of Scientific and Research Publications*.

[3] Kenkre, P. S., Pai, A., & Colaco, L. (2015). Real time intrusion detection and prevention system. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014* (pp. 405-411). Springer International Publishing.

[4] Shakshuki, E. M., Kang, N., & Sheltami, T. R. (2013). EAACK—a secure intrusion-detection system for MANETs. *IEEE Transactions on industrial electronics*, *60*(3), 1089-1098.

[5] Lin, W. C., Ke, S. W., & Tsai, C. F. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-based systems*, *78*, 13-21.

[6] Xiaodong, T., Wei, T., & Xianju, W. (2014). Application Research on Correlation between Intrusion Detection System and Vulnerability Scanner. *Network Security Technology & Application*, *8*, 076.

[7] Nadiammai, G. V., & Hemalatha, M. (2014). Effective approach toward Intrusion Detection System using data mining techniques. *Egyptian Informatics Journal*, *15*(1), 37-50.

[8] Faisal, M. A., Aung, Z., Williams, J. R., & Sanchez, A. (2015). Data-stream-based intrusion detection system for advanced metering infrastructure in smart grid: A feasibility study. *IEEE Systems Journal*, *9*(1), 31-44.

[9] Mukesh, K. G., Khanna, H. P., & Velvizhi, R. V. (2015). An anamoly based Intrusion Detection System for mobile ad-hoc networks using genetic algorithm based support vector machine. *Advances in Natural and Applied Sciences*, *9*(12), 40-45.

[10] Panda, M. and Patra, M.R., 2007. Network intrusion detection using naive bayes. International journal of computer science and network security, 7(12), pp.258-263.

[11] Syarif, I., Zaluska, E., Prugel-Bennett, A. and Wills, G., 2012. Application of bagging, boosting and stacking to intrusion detection. Machine Learning and Data Mining in Pattern Recognition, pp.593-602.

[12] Li, W., 2004. Using genetic algorithm for network intrusion detection. Proceedings of the United States Department of Energy Cyber Security Group, 1, pp.1-8.