Big Data Security Analysis in Network Intrusion Detection System

Muhammad Umer Farooq Alvi
Taiyuan University of Technology, China

Hao Xiaoli
Taiyuan University of Technology, China

Saad Abdul Rauf
Taiyuan University of Technology, China

ABSTRACT
This paper introduces Big data security analysis with the help of different techniques used in network intrusion detection system. The topic of how big data affects any intrusion detection system being used and how huge volume of the dataset, its specialized features that are heterogeneous in nature and what will happen if big data is processed at real time. Different attacks and intrusion detection methods such as intrusion detection and prevention systems (IDPS), signature-based detection (SD) and anomaly-based detection (AD) has been done. Challenges faced by intrusion detection systems (IDS), how they can be prevented and how machine learning, data mining techniques could be used in any general intrusion detection-based system has also been discussed. Also, how all the problem faced by IDPS can be solved by network simulator named NS-3.0. Its objectives, advantages, comparison with other networks and limitation have also been to be discussed. The recommendation is also given to improve faults. Also, results obtained after using NS-3 based svm classifier using KDD Cup 99 Dataset showed the accuracy of 99 percent.

Keywords
anomaly-based detection, big data security analysis, challenges, data mining, intrusion detection and prevention systems (IDPS) machine learning, network security, NS-3, signature-based detection, svm.

1. INTRODUCTION
The first thing should be known is about what is Big Data analytics (BDA), it deals with very complex process in which extremely varied large data sets are dealt in. It has the specialty to transfer enormous size and magnitude data very hurriedly efficiently and effectively (Oseku-Afful, 2016). To validate and manage this size of data is one of the rising problems in BDA these days. BDA has different types and forms such as internet blogs, social media sharing platforms, Facebook, Twitter, WordPress blogging. There must be some kind of rules to manage all this traffic and data, thus arising the use of big data security analysis (Dewa A., 2016) Any illegal activity or an act carried with the help of any computer technology device or network is considered as a cybercrime and violation of act of privacy and intrusion. There are many ways and classes of cybercrime which are also considered as terrorism such as try to attempt to authenticate access or pass any service or business network in an illegitimate manner (Moustafa, Creech, & Slay, 2017a) A number of solutions were presented and adopted after the increase in cybercrime and internet terrorism because it affected a great number of organizations all over the globe. Its solutions are first line security measures which are Firewall (A trusted network system designed especially for security purposes which monitors the traffic of any network upon basis of predetermined rules and regulations) , Cryptography (use of different coding techniques for data communication security in the presence of adversaries or third parties) and second-line security solution is the use of intrusion detection system (IDS) which is basically a software developed application or a computer-controlled device that continuously 24/7 monitors the whole system of an organization or a network for any kind of illegal intrusion, act, access, violation of companies policies authorization or malicious activity(Ata, 2017)Upon detection of the above mentioned the IDS acts immediately and effectively informing and alarming the client, administrator in control at the same time by its event management system known as SIEM. SIEM can distinguish between a false alarm and a real-time act or threat proficient (Abd & Hadi, 2018) IDS have the major advantage of monitoring internets protocols of many web servers such as Hypertext Transfer Protocol (HTTP), due to this reason IDS is known as protocol-based intrusion detection system (PIDS). IDS another specialization is that it can also monitor protocols which are only application-specific such as application protocol-based intrusion detection system (APIDS). It monitors the Structured Query Language (SQL) protocol database extensively and widely. (Moustafa, Creech, & Slay, 2017b)

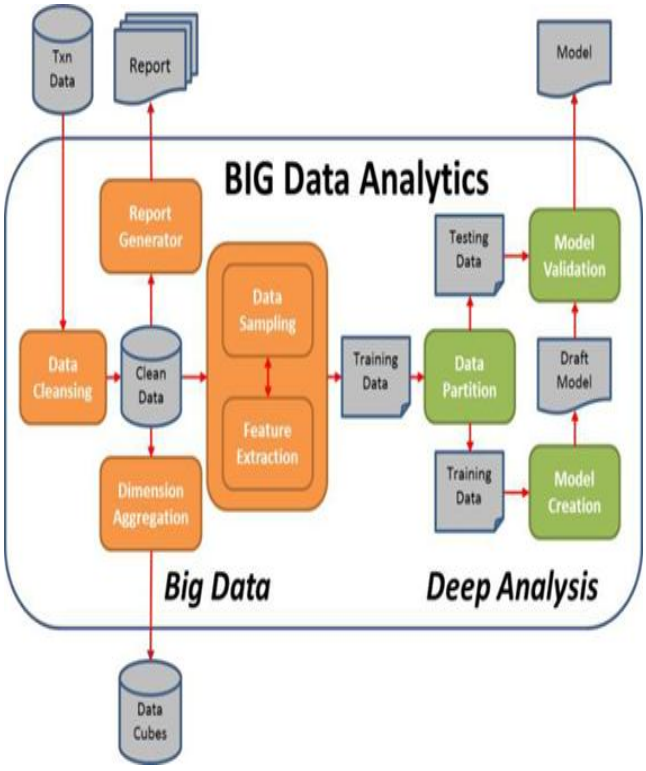

Fig 1: Big Data extensive overview

2. LITERATURE REVIEW

(Zeng, 2016) Has proposed a new technique named hierarchic IDS. It is a combination of any abnormal detection and misuse of deployed Power System. Hierarchic IDS has used Data mining algorithms for investigation and evaluation of data monitoring the detection methods applied by IDS based on data science theory include nerve network, support vector machine, immune, cluster analysis, data mining, cloud computing, and big data. Proposed implemented prototype is easily able to detect any kind of cyber-attacks with accurately ad quite a low rate of a false positive detection. Its specialty is ta it can handle any kind of Big Data and inform us any intrusion or attack. This technique also reduces the system dependence on any kind of experienced knowledge and artificiality in the intrusion detection system as it detects intrusion with the help of classic Data mining algorithms and techniques. (Moustafa et al., 2017a) Proposes an anomaly detection system which is not only effectively performs but also is lightweight to handle in terms of management. The prototype performance was evaluated with the datasets named NSL-KDD and UNSW-NB15. After application of prototype on NSL-KDD dataset its present accuracy came to be 97.8 % while its previous accuracy was 93.1 % and UNSW-NB15 dataset accuracy also increased from previous 84.1% to 93.9 %. The experimental results show that the implemented prototype selects and choses the best model which gives accurate results on the basis of given data arising the possibility of low false rate and has excellent rate of detection. (Wang, 2017) Has done a comparative analysis of techniques named intrusion detection systems, their prevention systems and different techniques of intrusion detection methods like anomaly-based detection and signature-based detection? He also discussed the problems, changes faced and posed by intrusion detection systems and how to prevent them. He also touched how big adept effect IDS and if the real-time stream is ongoing what must be done to prevent any intrusion. One important point of this study is that if more accuracy is to be achieved than such intrusion detection systems must be used which has multiple layers in it. (Hafsa & Jemili, 2018) has discussed Cybersecurity ventures and presented estimated the amount of eleven point five billion in 2019 damage done to organizations by different kinds of cyber-attacks. To put a stop to all this damage they proposed a system developed by them in software named Microsoft Azure Cloud which is a combination of Machine Learning library and Apache Spark Structured Streaming. The proposed prototype has the advantages of lower power usage during processing and enlarge, increased storage capabilities. The results obtained by them were up to 99.95% accurate.

3. METHODOLOGY

3.1 Frame work

Following is the flow diagram and frame work of adopted research, covering all the main points.

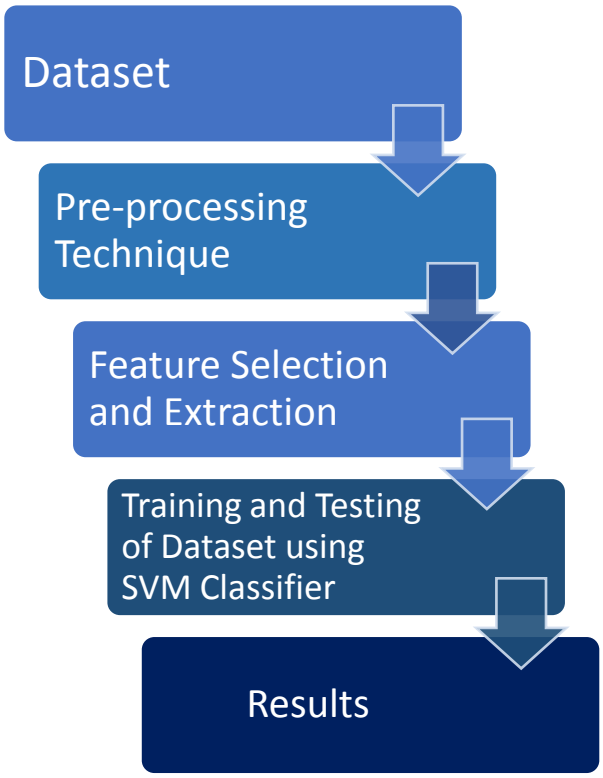

Fig 3: Adopted Research Technique- NS-3 based SVM-IDS Classifier

3.2 Big Data Security Analysis (BDSA) techniques

(Almansob, Jalil, & Lomte, 2017) For Big Data Security Analysis (BDSA) a number of well-known techniques are being used for cyber-attacks, cyber-terrorism and cyber-intrusion detection and prevention which are as follows,

•Antivirus Programs, •SIEM,

•File Integrity Monitoring •whitelisting (FIM),

•IDS

(Shackleford, 28 May 2016) All above mentioned techniques, algorithms and system have their advantages and effectiveness but they still are not strong enough to cover all types and ranges of cyber-attacks. Following are the reason for them in capabilities,

1. Huge volume of data.
2. Time consuming and difficult to manage.
3. Independent operating and working.
4. May miss important event considering at as false.
5. Real-time processing.
6. False detection and alarm alert.
7. Extensive data streaming

3.3 Intrusion Detection Systems

IDS can be hardware or software based on serving the propose of detecting and analyzing the any sign of intrusion and in it traffic of any network can also be analyzed both inbound and outbound at the same time. IDS detection is done by following undermentioned steps;

i. Data or System files are checked and compared for any malware or intrusion signals and signatures.

ii. Scanning is done for its rightful detection.

iii. System continuously monitored for any kind of cyber-attack. System configurators and settings for continuous monitoring and alert.

There is a problem with IDS and that is it can't detect any kind of incoming assaults, so for their blockage an intrusion prevention system is in demand and designed specifically. (Vyas, Meena, & Kumar, 2014)

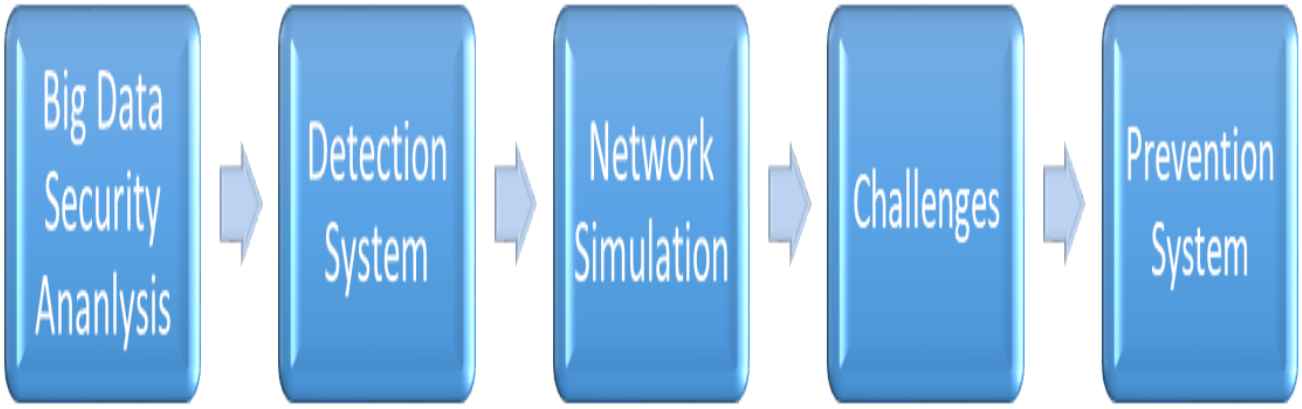

Fig 2: Adopted Research Frame work

3.3.1 IDS Classification

IDS data processing techniques are as follows.

[1] **Expert systems** – Such systems working is pre-defined because in it all the rules of attack identification have been fed. For example, Wisdom & Sense Computer Watch.

[2] **Colored Petri Nets** – it is generalization approach in which all attacks are represented graphically but still it is not used in commercial systems.

[3] **Machine learning** – an artificial technique in which input stream is fed and stored as a vector.

[4] **Signature analysis** – in it the system has all the knowledge and description of aback done is transformed in to appropriate audit trail and common text string matching mechanisms is used for detection. It has been deployed in Emerald eXpert-BSM commercial systems.

[5] **Statistical analysis approach** - with the help of number of variables over time system behavior is determined.

[6] **Data Mining** - is based on set of techniques which used the extraction of potentially useful from all data and is good at auditing of data.

[7] **State-transition analysis** –in it the attacker has been described with already set number of goals and many transitions.

[8] **User intention identification** –with the help of high-level tasks performed by user are being used to predict model's normal behavior.

[9] **Computer immunology Analogies** – in its model designed has a short sequences of system calls which has been made by the processes and helps in construction of a normal behavior model.

[10] **Neural Networks** – in it a system leers the inputs and output and their formed relationships and upon learning rate the new data prediction is done. In IDS based NN the systems learn what kind of attacks has been done on it before and then after learning it uses this information and predicts the kind of attack done. (Boutaba, 2013).

3.3.2 IDS Principles

Following are the main principles used in IDS.

a) **Signature-based Detection** – It detects intrusion or any attack with the help of patters such as the 1. sequence of bit or bytes present in the network traffic. Origin of this is from anti-virus software.

b) **Anomaly-based Detection** – also know these days as behavior-based detection. This method with the help of models such as computer systems, network and users, detects an intrusion event and raises alarm telling that the designed system is showing deviation from its normal behavior.

3.3.3 Machine Learning in Intrusion Detection System

Three types of Machine Learning classifiers such as single, hybrid and ensemble.

(i) **Fuzzy Logic** – it is used for purposes such as reasoning and range of value varies between 0 to 1. It is quite efficient in port scans and high resource consumption during IDS.

(ii) **Genetic Algorithms** – these algo make sure that the system or computes used have a natural ability to work with large and big datasets. Its performance is also better than other logics used.

(iii) **Self-Organizing Maps** – It is a type of neural network and uses unsupervised learning. Its specialty is to divide and map high dimension data in two-dimension array. Also reduces dimension and self-categorize all the inputs.

(iv) **K-Nearest Neighbor** – it classifies given samples by creating a K-NN classifier and uses the technique of finding distance between two different point

(v) **Support Vector Machine** – in its classification is done by hyper plan construction. It divides data into

two groups supports vectors and quadratic programming problem.

(vi) **Artificial Neural Networks** – it works like neurons of a human brain creating arterial perceptions. This technique is fast, flexible and efficient.

(vii) **Decision Trees** – It uses if then else rules. In it fits dataset attributes are chooses and then classification is done.

3.3.4 Data Mining in Intrusion Detection System

It is the process when from large datasets a number of peers are extracted with the help of artificial intelligence with database management. Data mining in IDS helps in removal of attack activity, identification of false alarm generation, log inform and bad activity.

3.3.5 Comparison of ID/PS Methods

Aspects	Methods	Advantages	Disadvantages
Detection method	anomaly	Has ability for detection of most new attacks on computer and networks. It has increased effectiveness, detection and efficiency classifying it It either classifies attack done as either normal and anomalous.	Difficulty to distinguish between boundaries of normal/abnormal behavior. To vulnerable have higher false positive alarms and events. It un-correctly detect an attack and then deliver it as correctly detected attack.
	Signature/ Misuse	It defines abnormal system behavior first and all other are considered as normal. Has low rate false generation. Reliable in known detection and specified attacks.	Due to poorly presence of constituted signature gives false alarms. Unable to recognize unknown attacks.
Audit source location	Host based	Less hardware required Effective cost range Easiest deployment in any system Ability to see low level activities Effective dealing with switched environments and encrypted data.	View of the network is very limited. As it is close to user so response to illegal tempering
	Network based	Quick and immediate response time. Has low false positive detection rate. It has monitoring network traffic which is present at the transport layer mostly missed by host-based systems.	It is far from individual hosts so unable for host implementation Can't decrypt data which is encrypted. Laborious evidence Removal
Data distribution modes	Distributed	It has the ability to investigate the security status after utilizing traffic information from various sources.	The data flow generates high network traffic overheads between host monitors and the director agent may.
	Central	In it all of the detection, monitoring, activities are controlled by a central console.	Attacker can easily destroy or modify data. Programs running on a system can be modified or disabled.
Time of detection	Non-real Time	Resource consumption is low Provides the high evidence of forensic data.	Unable to provide real time response Unable to prevent or mitigate damages.
	Real time	Good attack detection and prevention. Can fill the network inherent security gaps.	Unable to handle encrypted packets. Can't trace and responds attacks automatically.

3.3.6 Challenges of Intrusion Detection Systems

There are many challenges faced by IDS during Big data security data analysis, which are listed as follows.

i. IDS components have to communicate across sub-networks, network firewalls, and gateways. Due to this amount of hassle communication, it is sometimes not able to recognize all these different formats, which creates issues for the system.

ii. Large networks have effectively heavy and busy monitoring traffic. Due to this many attacks are able to bypass without being checked.

iii. Many networks have a limited time window for which the connection state is very difficult to maintain.

iv. Sometimes when the system is at real-time operation mode, IDS is unable to inspect each packet thus causing packet loss and increases inefficiency.

v. High traffic load reduces system power.

vi. A number of false positives are quite high due to a number of errors.

4. NETWORK SIMULATOR 3.0

After a comprehensive analysis, It came to the conclusion that the network simulator 3.0 would be preferable for covering all the challenges and technical deficiency.. NS-3 is a discrete-event network simulator which is mostly for Internet systems and networks. It is faster because of its lowest computation time is taken as compared to other software.

4.1 Objectives of NS-3

The following goals and objectives of NS-3 which they considered important:

1. To adopt community-oriented and open source development practices for all user.
2. To distribute free and open source software.
3. To leverage and permit the inclusion of free of cost and open source networking software.
4. To architect a well scalable, extensive, modular, emulator, clear and a well-documented simulator.
5. Core models should be well tested and validated.
6. To develop canonical simulation-based experiments.
7. To make a simulator for its use in networking courseware.

4.1.1 Comparison with other Network simulators

1. **NS2** – although this simulator has a wide range of support as it is not actively maintained, so there are many bugs that have been occurring in it. It also utilizes a large amount of CPU percent.

2. **OPNET**- which specializes in fields such as telecommunication designing, product manufacturing, and marketing for access and inter-office networks will also do, but it has proprietary issues.

3. NETSIM - creates a national cyber exercise. It is also considered a very fast and functional simulator. It is intended for its use in different kinds of defense

systems. It will support computer-based collaborative work and has a very effective built-in integrated debugging environment.

4. **OMNeT**++ is a C++ based discrete event simulator which is used for modeling of communication networks, distributed systems, parallel systems, and multiprocessors.

4.1.2 Advantages of NS-3

1. NS-3 provides extensive features not available in ns-2.
2. It has code executable environment.
3. It allows all users to run real-time implementation of any kind of code in the simulator environment.
4. It also provides a lower base level of abstraction.
5. It allows the written code to align better with how real systems work.
6. It also supports multiple interfaces on nodes correctly with different kinds and working.
7. It is actively maintained.
8. Has an active and very responsive user's mailing list.

4.1.3 Limitations

1. It only works on single-interface nodes.
2. When NS-3 attempts to run single nodes on a multi-interface node, a program error is caused.
3. Does not support Cygwin and OS X PPC.
4. Does not support non-Linux stacks.
5. Does not support socket API callbacks.

4.1.4 Accuracy results

The essential purpose of testing is to determine that a piece of software behaves "correctly." For NS-3 this means that the simulation of something should faithfully represent some physical entity or process to a specified accuracy and precision.

1. It provides tools to allow for both model validation and testing, and encourages the publication of validation results.
2. It allows for setting up and running test environments over multiple systems (buildbot) and provides classes to encourage clean tests to verify the operation of the system over the expected "domain of applicability" and "range of accuracy.
3. It provides tools for automating the process used to validate and verify the code in nightly test suites to help quickly identify possible regression.

4.2 SVM NS-3 Based IDS Classifier

4.2.1 Data set

KDD99, DARPA1999, and DARPA1998 are the data sets, most used for classification tasks for IDS. KDD99 is the most used data set for all purposes, as it is quite efficient in terms of attack detection but there is a number of problems with the dataset DARPA because of its late, false and un-realistic behavior when it is being attacked by an intruder. NSL-KDD dataset has 41 features, class label named normal and four attacks named DoS, Probe, U2R, and R2L. The dataset that are selected for IDS classification purposes is KDD99.

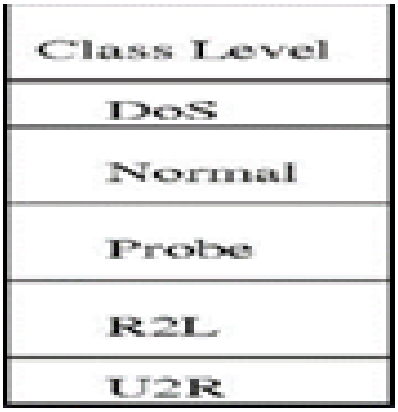

Fig 4: KDD99 Classes

No.	Network attributes	No.	Network attributes	No.	Network attributes
1	duration	15	su_attempted	29	same_srv_rate
2	protocol_type	16	num_root	30	diff_srv_rate
3	service	17	num_file_creations	31	srv_diff_host_rate
4	flag	18	num_shells	32	dst_host_count
5	src_bytes	19	num_access_files	33	dst_host_srv_count
6	dst_bytes	20	num_outbound_cmds	34	dst_host_same_srv_rate
7	land	21	is_host_login	35	dst_host_diff_srv_rate
8	wrong_fragment	22	is_guest_login	36	dst_host_same_src_port_rate
9	urgent	23	count	37	dst_host_srv_diff_host_rate
10	hot	24	srv_count	38	dst_host_serror_rate
11	num_failed_logins	25	serror_rate	39	dst_host_srv_serror_rate
12	logged_in	26	srv_serror_rate	40	dst_host_rerror_rate
13	num_compromised	27	rerror_rate	41	dst_host_srv_rerror_rate
14	root_shell	28	srv_rerror_rate		

Fig 5: KDD99 Features

4.2.2 Steps

Following number of steps are adopted to classify and identify KDD99 5 classes in network simulator NS-3.

1. The dataset was loaded in NS-3 simulator by reading the file named kddcup99_csv.
2. Preprocessing techniques were applied on it.
3. Among all extracted features best features were selected.
4. These features were fed to NS-3 based SVM classifier.
5. Final results were obtained.

5. RESULTS

Following are the results obtained after running written script for SVM-IDS classifier on NS-3

5.1 Feature Selection

1. Feature named dst_host_same_src_port_rate has very less effect on the intrusion type but when value greater than equal to 1 "probe" and "r2l".
2. Feature named "flag" is a strong predictor. Thus, when flag is equal to "REG" and "S0" and it classifies "dos".
3. Feature named "duration" If its value approaches

30000, the predictor is high and can classify ' probe.

4. Feature named "protocol-type", "tcp" acts as "DOS" intrusion type. Thus, it is also a strong predictor and classifies "dos" type.
5. Feature named serror_rate and srv_serror_rate, gives no clear identification and classification of any type.
6. Feature named serror_rate and srv_serror_rate value is either 0 or 1 it classifies "dos" and when feature named serror_rate value is between 0.25 to 0.5, "probe" is classified.

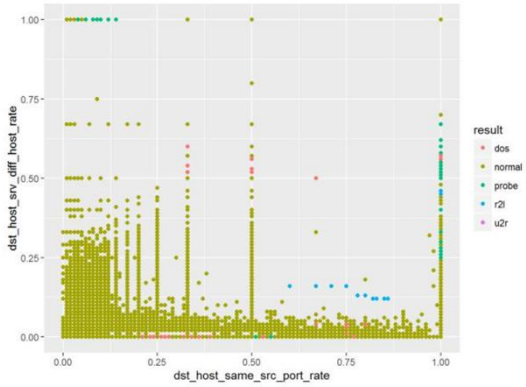

Fig 6: dst_host_same_src_port_rate vs dst_host_srv_diff_host_rate, color

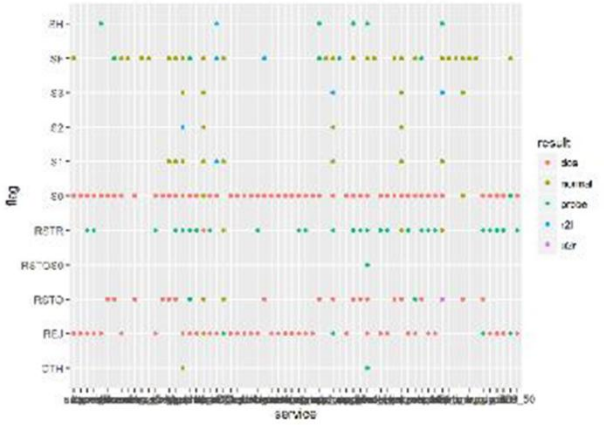

Fig 7: service, vs flag

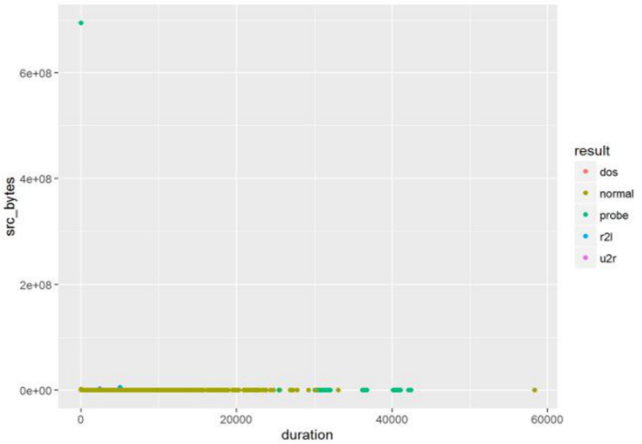

Fig 8: duration vs src_bytes

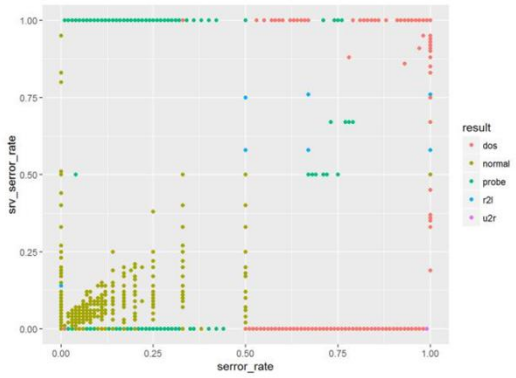

Fig 9: serror_rate vs srv_serror_rate

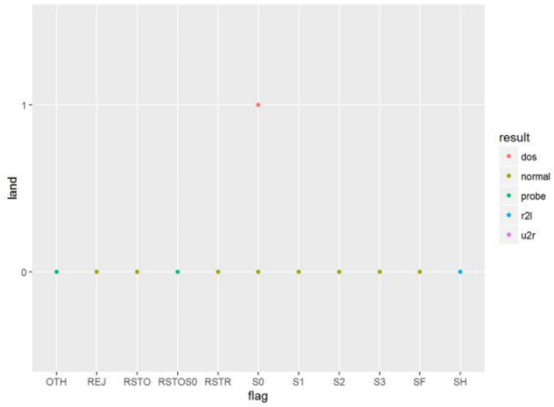

Fig 10: flag vs land

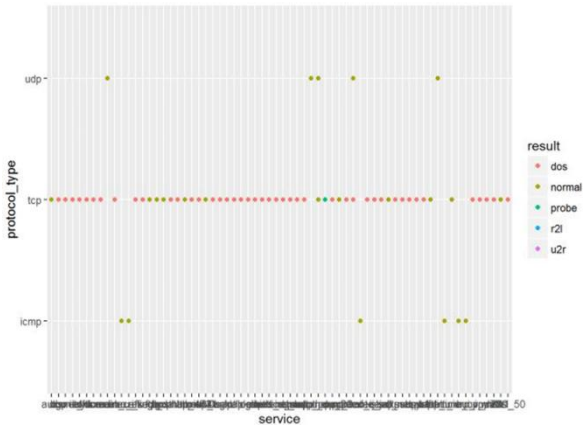

Fig 11: service vs protocol type

```
49404 samples
   25 predictor
    5 classes: 'dos', 'normal', 'probe', 'r2l', 'u2r'
```

Fig 12: NS-3 based SVM classifier

pred	dos	normal	probe	r2l	u2r
dos	352278	14	9	3	0
normal	23	87509	30	79	41
probe	11	13	3657	2	0
r2l	0	13	0	928	1
u2r	0	1	0	1	4

Figure 13- predictor vs testing result (Number of samples)

pred	dos	normal	probe	r2l	u2r
dos	99.99	0.00	0.00	0.00	0.00
normal	0.03	99.80	0.03	0.09	0.05
probe	0.30	0.35	99.29	0.05	0.00
r2l	0.00	1.38	0.00	98.51	0.11
u2r	0.00	16.67	0.00	16.67	66.67

Figure 14- predictor vs testing result (Accuracy Result)

6. CONCLUSION

Big Data analytics (BDA) usually shifts through heterogeneous systems with an extensive huge amount of data with quick, efficient and effective speed. A system which has multiple types of IDPS technologies are considered accurate, comprehensive performance and increased efficiency. The designed IDS and IPS must always recognize all data formats, communication protocols, network routes, and critical paths. In this study, the proposed technique that simulator NS-3 must be used to overcome listed challenges and errors. Thus, whatever BDA security analysis is being used it must be compatible and strong enough to knock out all types of attacks without putting the IDS at risk. The accuracy of NS-3 based SVM-IDS classifier is about 99 percent accurate.

7. FUTURE RECOMMENDATIONS

Following are the recommendations that should facilitate more efficient and effective intrusion detection and prevention system.

1. All IDPS systems must be secured appropriately as these systems are targeted by attackers on very large scale.
2. For prevention of gain of system sensitive information, the intrusion detection and prevention system must have a strong host configuration and no existing vulnerabilities.

8. REFERENCES

[1] Abd, A., & Hadi, A. (2018). Performance Analysis of Big Data Intrusion Detection System over Random Forest Algorithm. International Journal of Applied Engineering Research ISSN, 13(2), 973–4562. Retrieved from http://www.ripublication.com

[2] Almansob, S. M., Jalil, A. A., & Lomte, D. S. S. (2017). The Use of K-NN and Bees Algorithm for Big Data Intrusion Detection System. IOSR Journal of Computer Engineering, 19(01), 08–12. https://doi.org/10.9790/0661-1901040812

[3] Ata, B. I. G. D. (2017). R Eal -T Ime I Ntrusion D Etection S Ystem F or. 8(1), 1–20.

[4] Boutaba, R. (2013). Intrusion Detection. Intrusion Detection Networks, (June), 21–37. https://doi.org/10.1201/b16048-5

[5] Dewa, Z., & A., L. (2016). Data Mining and Intrusion Detection Systems. International Journal of

[6] Advanced Computer Science and Applications, 7(1), 62–71. https://doi.org/10.14569/ijacsa.2016.070109

[7] Hafsa, M., & Jemili, F. (2018). Comparative Study between Big Data Analysis Techniques in Intrusion Detection. Big Data and Cognitive Computing, 3(1), 1. https://doi.org/10.3390/bdcc3010001

[8] Moustafa, N., Creech, G., & Slay, J. (2017a). Data Analytics and Decision Support for Cybersecurity. https://doi.org/10.1007/978-3-319-59439-2

[9] Moustafa, N., Creech, G., & Slay, J. (2017b). Data Analytics and Decision Support for Cybersecurity. https://doi.org/10.1007/978-3-319-59439-2

[10] Vyas, G., Meena, S., & Kumar, P. (2014). Intrusion Detection Systems: A Modern Investigation. (11), 4–7.

[11] Wang, L. (2017). Big Data in Intrusion Detection Systems and Intrusion Prevention Systems. Journal of Computer Networks, 4(1), 48–55. https://doi.org/10.12691/jcn-4-1-5.

[12] Zeng, S. (2016). An Intrusion Detection System Based on Big Data for Power System. (Isaeece), 322– https://doi.org/10.2991/isaeece-16.2016.62