

Detecting and Analysing Network Logs Using Machine Learning Techniques

Akhila Anilkumar¹; Alona Shibu²; Meera Anna Varghese³; Priya P Sajan⁴; A.L. Sreedeeep⁵

¹Bachelor in Technology-IT with Specialization in Cyber Security and Forensics, Noorul Islam Centre for Higher Education.

¹aakhila958@gmail.com

²Bachelor in Technology-IT with Specialization in Cyber Security and Forensics, Noorul Islam Centre for Higher Education.

²alonas734@gmail.com

³Bachelor in Technology-IT with Specialization in Cyber Security and Forensics, Noorul Islam Centre for Higher Education.

³meeraannavarghese@gmail.com

⁴Project Engineer, C-DAC Thiruvananthapuram.

⁴priyasajan@cdac.in

⁵Project Engineer, C-DAC Thiruvananthapuram

⁵sreedeeep@cdac.in

Abstract

A Network rhetorical investigation needs the capturing, recording and analysing of network proof and therefore the audit trails. The end result of such investigations could deliver security audit, security data to harden a system or proof for legal functions. One of the most common method of attack involve sending large amount of request to sites or server and server will be unable to handle and sites will be offline for many days. The aim is at determining the verity by analysing the network activity logs. Some machine learning algorithms are deployed to find various attack like DoS UR2, R2L and probe type. The mooted machine learning algorithms are SVM, Random Forest, Decision Tree, Logistic Regression, Naïve Bayes and K-nearest neighbour. Multiple network logs are collected and analysed using Wireshark. Log analysis is then availing to find the quandary in period of time and fine-tune it before it engenders ruin. A mechanism designed to aggregate and analyses these logs to have a clear overview of what's transpiring across the network to determine various attacks. And finally, the effectiveness and accuracy of various machine learning algorithms on log datasets having particular features are evaluated. Thus, indemnity from attackers could be enhanced by providing dependability, solidity and surety.

Key-words: Network Logs, Wireshark, Log Analysis, Machine Learning Techniques.

1. Introduction

An intrusion detection system merely monitors network traffic and should alert the network administrator of any uncommon activity. it's terribly almost like an intrusion detection system merely monitors network traffic and should alert the network administrator of any uncommon activity. it's terribly almost kind of a house alarm which is during a position to sound AN alarm if an intruder makes an attempt to interrupt into a window or a door. as an example, if a hacker makes an attempt to know access to your pc or network, the intrusion detection system can immediately advise the network administrator of the tried security contravention. Logs act as a red flag. With security log analysis, you'll be able to hunt suspicious activities and created thresholds, rules, and parameters to guard your system from similar threats within the long run. With log analysis, you're even able to assist in obstruction your attackers by their science address. Log information analysis systems will provide you with a warning whenever they discover anomalies so you'll be able to quickly intervene and eliminate the threat. a house alarm which is during a position to sound AN alarm if an intruder makes an attempt to interrupt into a window or a door. as an example, if a hacker makes an attempt to know access to your pc or network, the intrusion detection system can immediately advise the network administrator of the tried security contravention. Logs act as a red flag. With security log analysis, you'll be able to hunt suspicious activities and created thresholds, rules, and parameters to guard your system from similar threats within the long run. With log analysis, you're even able to assist in obstruction your attackers by their science address. Log information analysis systems will provide you with a warning whenever they discover anomalies so you'll be able to quickly intervene and eliminate the threat.

They use computing and machine learning to identify patterns and behaviours which can have otherwise flown below the radar. what's more, logs are very helpful in cyber forensics. just just just in case of AN investigation, rhetorical log analysis will give the time and place of each event that happened in your network or system

2. Background

Machine Learning

Machine learning could even be a multidisciplinary approach at the start employed in supervised learning to make analytical models. It plays a big facet in an exceedingly broad scope of great applications like image recognition, processing, delicate systems and image recognition. This

approach seems appropriate to resolve phishing page detection, as a result of this drawback could even be reborn into a task of classification. millilitre techniques could even be wont to develop models to sight phishing activities supported categorizing previous websites then these models could even be integrated into the browser

INTRUSION DETECTION SYSTEM: An Intrusion Detection System (IDS) could be a system that monitors network traffic for suspicious activity and problems alerts once such activity is discovered. it's a software application that scans a network or a system for harmful activity or policy contraventioning. Any malicious venture or violation is typically reportable either to Associate in Nursing administrator or collected centrally employing a security data and event management (SIEM) system. A SIEM framework coordinates yields from numerous sources and uses caution sifting strategies to separate malignant movement from bogus alerts.

3. Types of Attack

DDoS Attack

Distributed denial of service (DDoS) attack is once associate aggressor, or attackers, decide to build it impossible for a service to be delivered. this could be achieved by thwarting access to only about anything: servers, devices, services, networks, implementations, and even specific transactions at intervals applications. in an exceedingly DoS attack, it's one system that's causation the malicious information or requests; a DDoS attack comes from multiple systems. Generally, these attacks work by drowning a system with requests for information. this might be causation a web server such plenty of requests to serve a page that it crashes beneath the demand, or it would be an info being hit with a high volume of queries. the outcome's available web arrangement of estimation, focal preparing unit and RAM ability gets inundated.

Denial-of-service Attack (DoS attack)

Denial-of-service attack (DoS attack) could also be a cyber-attack within which the culprit seeks to form a machine or network resource unprocurable to its supposed users by briefly or indefinitely disrupting services of a bunch connected to internet. Denial of service is usually accomplished by flooding the targeted machine or resource with superfluous requests in an endeavour to overload systems and forestall some or all legitimate requests from being consummated.

Remote to User (R2L)

Remote to user (R2L) is one kind of network attacks, within which associate persona non grata sends set of packets to a special laptop or server over a network wherever he/she doesn't have permission to access as a neighbourhood user. Remote to native attack (r2l) has been wide glorious to be launched by associate aggressor to understand unauthorized access to a victim machine within the entire network.

User to Root Attack (U2R)

User to root attack (u2r) is typically launched for illicitly getting the root's privileges once de jure accessing a neighbourhood machine. our major attack class like denial of service, probe, user-to root, and remote-to-local. This paper targeted on user-to-root attack, that the aggressor tries to access traditional user account and gains root access data of the system. The U2R attacks results in many vulnerabilities like sniffing parole, a wordbook attack and social engineering attacks.

Probe Attack

Probe Attack have associate access to entire network data before introducing associate attack. Examples: ipsweep, Nmap Probe is an endeavour to understand access to a laptop and its files through a superb or probable liability within the pc system. Network Probes don't seem to be a directly threat. However, they're doing indicate that somebody is casing your system for potential entry points for attack. It's a network monitor that analyses protocols and network traffic (in real-time). The definition of a network probe will vary counting on the network management application your victimisation.

Smurf Attack

A Smurf attack could even be a selection of a distributed denial of service (DDoS) attack that renders laptop networks inoperable. The Smurf program accomplishes this by exploiting vulnerabilities of internet Protocol (IP) and web management Message Protocols (ICMP).

Perl Attack

Preventing Cross website Scripting Attack the cross-site scripting attack is one altogether the foremost common, nevertheless unnoted, security issues facing internet developers nowadays. a web site is powerless in the event that it shows client submitted content on trust for malevolent content labels.

Xsnoop

Snooping attacks involve associate persona non grata taking note of traffic between 2 machines on your network. If traffic includes passing unencrypted passwords, associate unauthorized individual will doubtless access your network and browse tip.

XTerm

Denial of service attack Xterm features a feature to vary the title of the xterm window by causation one altogether the escape codes of the xterm. (Linux: man console_codes).

SATAN

Satan can dutifully scan your network and report back all the potential weaknesses that it finds—that is its job. it'll even tell you the way those weaknesses might be exploitable. it'll not fix any issues or keep unwanted guests out—that is your job. No program are getting to be a substitute for associate smart Security Administrator.

Neptune Attack

Neptune attack and Flash Crowd are 2 typical threats to internet servers. These 2 anomalies have several identical options that build them troublesome to tell apart.

Nmap

Nmap could even be a network clerk that has emerged collectively of the foremost a la mode, free network discovery tools on the market. Nmap is currently one altogether the core tools employed

by network administrators to map their networks. The program is getting to be accustomed realize live hosts on a network, perform port scanning, ping sweeps, OS detection, and version detection.

BackDoor

A backdoor attack could be a kind of malware that provides unauthorized access to a web site. Cybercriminals introduce the malware through unstable marks of passage, as outdated modules or info fields. Once they enter through the rear door, they need access to all or any or any or any your company's information, in conjunction with customers' personal classifiable data (PII).

Multi-hop Routing

Multihop routing could be a kind of communication in radio networks within which network coverage space is larger than radio vary of single nodes. Therefore, to understand some destination a node will use alternative nodes as relays. Since the transceiver is that the most supply of power consumption during a radio node and long-distance transmission needs high power, in some cases multi-hop routing are getting to be additional energy economical than single-hop routing.

Worm

A laptop worm could be a kind of malware that spreads copies of itself from laptop to laptop. A worm will replicate itself with none human interaction, and it doesn't get to attach itself to a package program so on cause injury.

Worms will modify and delete files, which they go to even inject further malicious package onto a laptop. Typically, a laptop worm's purpose is barely to form copies of itself over and over — depleting system resources, like disc drive house or system of measurement, by overloading a shared network. additionally, to wreaking disturbance on a computer's resources, worms can even steal information, install a backdoor, and permit a hacker to understand management over a laptop and its system settings.

Buffer Overflow

A buffer overflow (or buffer overrun) happens once the degree of knowledge exceeds the storage capability of the memory buffer. As a result, the program making an attempt to write down

the knowledge to the buffer overwrites adjacent memory locations. Buffer overflows will have an impression on every kind of package. they typically result from unshapely inputs or failure to apportion enough house for the buffer. If the dealings overwrite possible code, it'll cause the program to behave erratically and generate incorrect results, operation errors, or crashes.

SQL Injection (SQLi)

SQL Injection (SQLi) could be a kind of associate injection attack that creates it potential to execute malicious SQL statements. These statements management an info server behind a web application. Aggressors will utilize SQL Injection weaknesses to sidestep application safety efforts. they'll go around authentication and authorization of internet a web an online} page or web application and retrieve the content of the whole SQL info.

SAINT

SAINT (Security Administrator's Integrated Network Tool) is laptop package used for scanning laptop networks for security vulnerabilities, and exploiting found vulnerabilities. The scanning method begins by detection all live targets among the given target list or vary. the chosen scanning policy then determines that core probes area unit run against every target. Results from the probes area unit utilized by the reasoning engine to schedule further probes and to infer vulnerabilities and alternative data supported rule sets. Final scan results area unit then keep within the back-end info to support information analysis and coverage through either the browser interface, statement Interface (CLI) or accessed via the appliance programming interface (API).

Raffaele Bolla and Roberto Bruschi in their work A high-end Linux based Open Router for IP QoS networks: tuning and performance analysis with internal (profiling) and external measurement tools of the packet forwarding capabilities [1] tries to supply a commitment by announcing the consequences of a profound action of advancement and testing acknowledged on a PC Open Router architecture supported Linux software. the primary goals are the exhibition assessment (regarding packet sending) of a sophisticated OR, both with outer (throughput) and interior (profiling) estimations.

Arpit Shah and Nilesh Kakade in their work A Survey: Investigation for Apache Log Pre-processing in Web Usage Mining [2] The Principal Procedure is 3 Phases, Information Cleaning,

User Recognizable evidence and Session Distinguishing evidence. The execution of the data cleaning system of blog data using Web use mining Procedure.

Dimitrios Xanthidis and Dr David Nicholas in their work An Investigation of the use or not of blog analysis by online businesses [3] The Principle Procedure is 3 Phases, Information Cleaning, User Recognizable evidence and Session Distinguishing Evidence. are execution of the knowledge cleaning procedure of blog information utilizing Web utilization mining Procedure.

4. Proposed Work

Methodologies

- WIRESHARK- Used for analysing network traffic and logs
- MACHINE LEARNING ALGORITHMS - Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, K-Nearest Neighbours and Decision Tree used to acquire exactness and accuracy.

Support Vector Machine: SVM is enforced unambiguously in comparison to different Machine Learning algorithms. An SVM coaching rule builds a model that assigns new examples to 1 class or the opposite, creating it a non-probabilistic binary linear classifier. SVM may be a supervised Learning rule, that is employed for Classification moreover as Regression issues. In any case, principally, it's utilized for Classification issues in Machine Learning. Additionally, to activity linear classification, SVMs will expeditiously perform a non-linear classification moreover employing a trick or parameter known as Kernel, that implicitly maps their inputs into high-dimensional feature areas. However, it's principally employed in classification issues.

Random Forest: Random Forest: Random Forest may be a supervised learning algorithmic rule. The "backwoods" it fabricates, is A gathering of call trees, now and again prepared with the "packing" philosophy. the last arrangement of the material system is that a combination of learning models will build the outcome. place just: irregular timberland assembles numerous call trees and unions them along to ask a ton of right and stable expectation. One enormous benefit of irregular woodland is that it is frequently utilized for every grouping and relapse issues, that type the heft of current AI framework.

Logistic Regression: Calculated relapse is one in all the preeminent in style Machine Learning calculations, that comes under the administered Learning strategy. it's utilized for anticipating the particular variable amount utilizing a given arrangement of independent factors. calculated relapse

predicts the yield of an absolute variable amount. Consequently, the outcome ought to be an absolute or unmistakable cost. It is frequently either confirmed or no, 0 or 1, valid or False, and so forth anyway as opposed to giving the exact cost as nothing and one, it offers the probabilistic qualities that lie somewhere in the range of nothing and one. calculated Regression is far similar to the relapse with the exception of that anyway they're utilized. relapse is utilized for discovering Regression issues, while strategic relapse is utilized for discovering the order issues. In strategic relapse, as opposed to fitting a relapse bend, we tend to coordinate an "S" framed calculated play out, that predicts 2 most qualities (0 or 1).

Naïve Bayes: Naive Bayes may be a probabilistic machine learning algorithmic rule supported the Bayes Theorem, utilized in a good type of classification tasks. during this post, you may gain a transparent and complete understanding of the Naive Thomas Bayes algorithmic rule and every one necessary idea so there's no space for doubts or gap in understanding.

Decision Tree: A tree might be a flowchart-like tree structure any place an indoor hub addresses highlight (or property), the branch addresses a decision rule, and each leaf hub addresses the end product. The top hub in an extremely call tree is perceived in light of the fact that the root hub. It figures out how to segment on the possibility of the quality cost. It parcels the tree in algorithmically way choice recursive dividing. This flowchart-like construction helps you in higher psychological cycle. It's perception kind of a multidimensional language outline that essentially impersonates the human level reasoning. that is the reason call trees region unit easy to know and decipher.

K-Nearest Neighbor (KNN): K-Nearest Neighbor is one in the very best Machine Learning calculations upheld managed Learning method. K-NN algorithmic standard expects the similitude between the new case/information and available cases and spot the new case into the class that is generally similar to the available classes. K-NN algorithmic principle stores all the available information and groups a substitution datum upheld the comparability. this infers once new information appears to be then it are frequently essentially characterized into a well suite class by exploitation K-NN algorithmic rule. K-NN algorithmic standard are regularly utilized for Regression likewise concerning Classification anyway basically it's utilized for the Classification issues. K-NN might be a non-parametric algorithmic principle, which proposes it doesn't construct any supposition on fundamental knowledge. It is furthermore alluded to as a languid student algorithmic guideline because of it doesn't gain from the instructing set continuously rather it stores the dataset and at the hour of grouping, it plays out a movement on the dataset.

WIRESHARK: It is an open-source packet instrument, that is employed for education, analysis, computer code development, communication protocol development, and network troubleshooting. it's accustomed track the packets in order that everyone is filtered to satisfy our specific wants. it's usually known as as a person, network protocol analyzer, and network analyzer. The network traffic is been monitored and analyse the network logs whether or not any style of attacks found and sight them. The planned system is split into 2 modules:

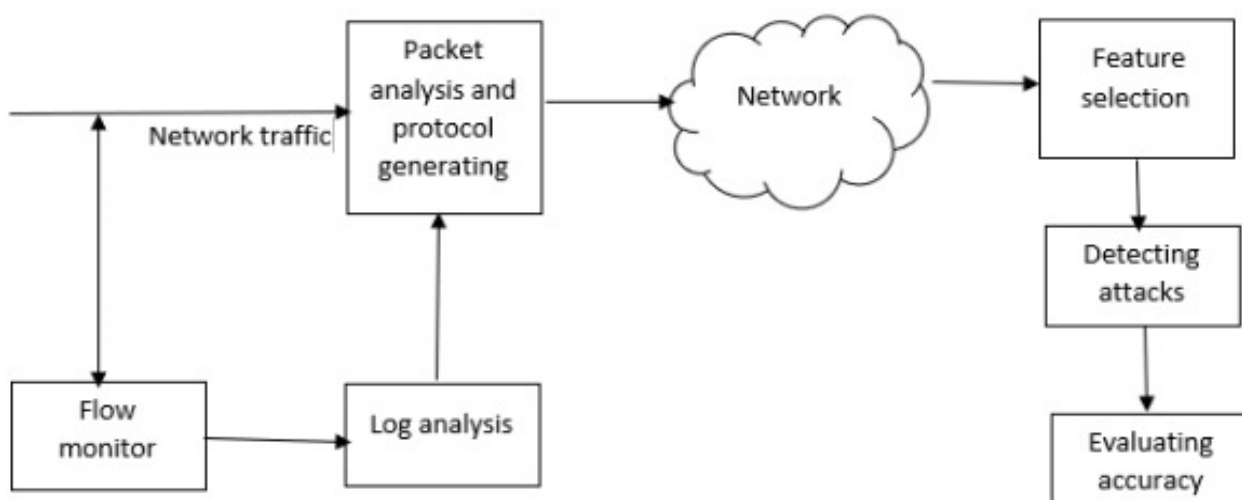
Network and Log Analysis using Wireshark

1. Log Analysis using Machine Learning

Phase 1: In the first phase, Wireshark tool which is used to analyse and track the packet, and find any attack has been occurred. After finding the attack from the given logs, the logs are converted to dataset for further analysis.

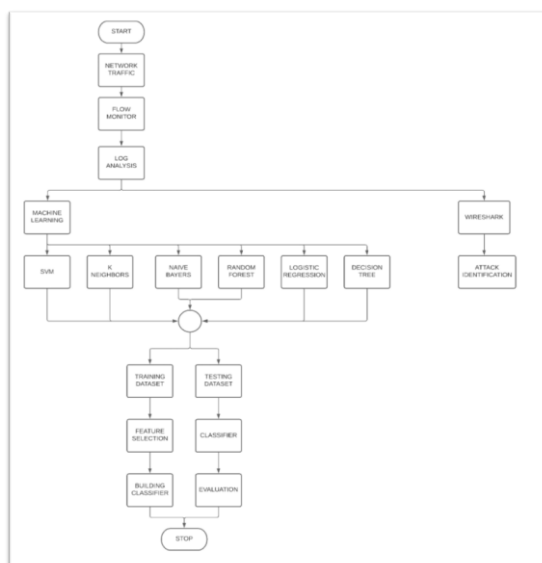
Phase 2: The given log dataset is used to predict the accuracy rate using the six machine learning algorithms Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, K-Nearest Neighbour, Decision Tree. Then the dataset is tested and trained using the algorithm and also extract the features of the given logs and build the classifier Using these six algorithms it shows which machine learning algorithm predict maximum accuracy rate and compare them with other classifiers using different static methods. It is conjointly employed by network security engineers to look at security issues.

Fig - Block Diagram



The model is shown in the block diagram where the network traffic is analysed and flow is monitored to get the log analysis done. The traffic packets are analysed to determine their protocol and include it for analysis. Once log analysis is done the data is sent to the network where features are selected the using them attacks are detected. The attack results are used to evaluate the accuracy which shows the performance level of the process.

Fig - Flowchart



The flowchart shows the data stream where the interaction begins from analysing network traffic to monitor the flow and set up log analysis of the network. The log analysis is carried out in Wireshark to detect types of attack that occurred in the log. This log is converted to dataset for analysis by machine learning algorithms. The algorithms are namely svm, logistic regression, random forest, KNN, naive Bayes and decision tree. The dataset is trained with respect to each algorithm and tested Accordingly to acquire details on the attack in the log.

5. DATASET

A Data set may be a set or assortment of knowledge. This set is generally conferred in an exceedingly tabular pattern. a knowledge set consists of roughly 2 parts.

DATASETS USED ARE: ftplogs, ddos, icmp, security logs, network traffic logs, network traffic practice logs, noobs lumberjacks.

- Ftp logs: The FTP log contains a record of all FTP associations anyway avoids any associations made by means of SFTP/SSH.

- Ddos logs: DDoS attacks are illegal under the Computer Fraud and Abuse Act.
- Icmp logs: ICMP is actually an integral part of IP.
- security logs: A security log is employed to trace security-related info on a computing system.
- network logs: an organization log is regularly a document that contains a record of occasions that happened in the application.
- network traffic exercise logs

Features extracted from the datasets used:

- CFSE: - Correlation based feature selection subset evaluator.
- CSE: - Consistency subset evaluator.
- Shared: - Shows features that are shared by both CSE and CFSE. Give a relatively poor intrusion detection rate.
- Combined: - Shows all the features in CFSE and CSE. should give a very good intrusion detection rate.
- Proposed: - Shows the neglected features.

Table 3.1: Feature Extraction

FEATURE SET	SELECTED FEATURES	NO. OF FEATURES
CFSE	Service, dst_bytes, logged_in, root_shell, count, dsthost_count, srv_ddiff_host_rate, dst_host_srv_diff_host_rate	8
CSE	Service, src_bytes, dst_bytes, logged_in, count, dst_host_srv_count, dst_host_diff_srv_rate, dst_host_rerror_rate	8
Combined	Service, dst_bytes, logged_in, root_shell, count, srv_diff_host_rate, dst_host_count, dst_host_srv_diff_host_rate, src_bytes, dst_host_srv_count, dst_host_diff_srv_rate, dst_host_rerroe_rate	12
Shared	Service, dst_bytes, logged_in, count	4
Proposed	Service, dst_bytes, logged_in, count, dst_host_count*, root_shell*, dst_host_rerror_rate*	7

6. Result and Discussions

Log analysis is that the method of reviewing and understanding logs to get valuable insights.

The first reason for activity log analysis is additionally a number of the foremost necessary reasons to perform work itself, particularly troubleshooting issues.

Table - Accuracy and Protocol of Classifiers

SL. NO	CLASSIFIERS	PROTOCOL	ACCURACY
1	RANDOM FOREST	ICMP	99.6
2	SUPPORT VECTOR MACHINE	ICMP	99.6
3	K-NEAREST NEIGHBOUR	ICMP	99.6
4	LOGISTIC REGRESSION	ICMP	99.53
5	DECISION TREE	ICMP	99.53
6	NAVIE BAYES	ICMP	15.9

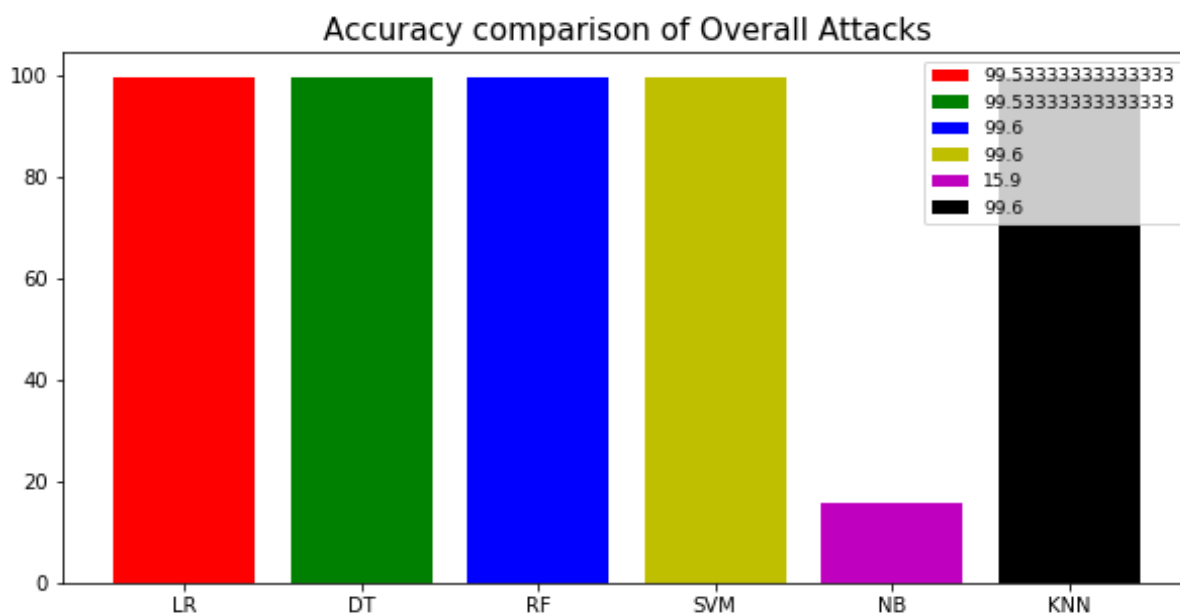


Fig - Graph for Overall Attack Using Classifiers

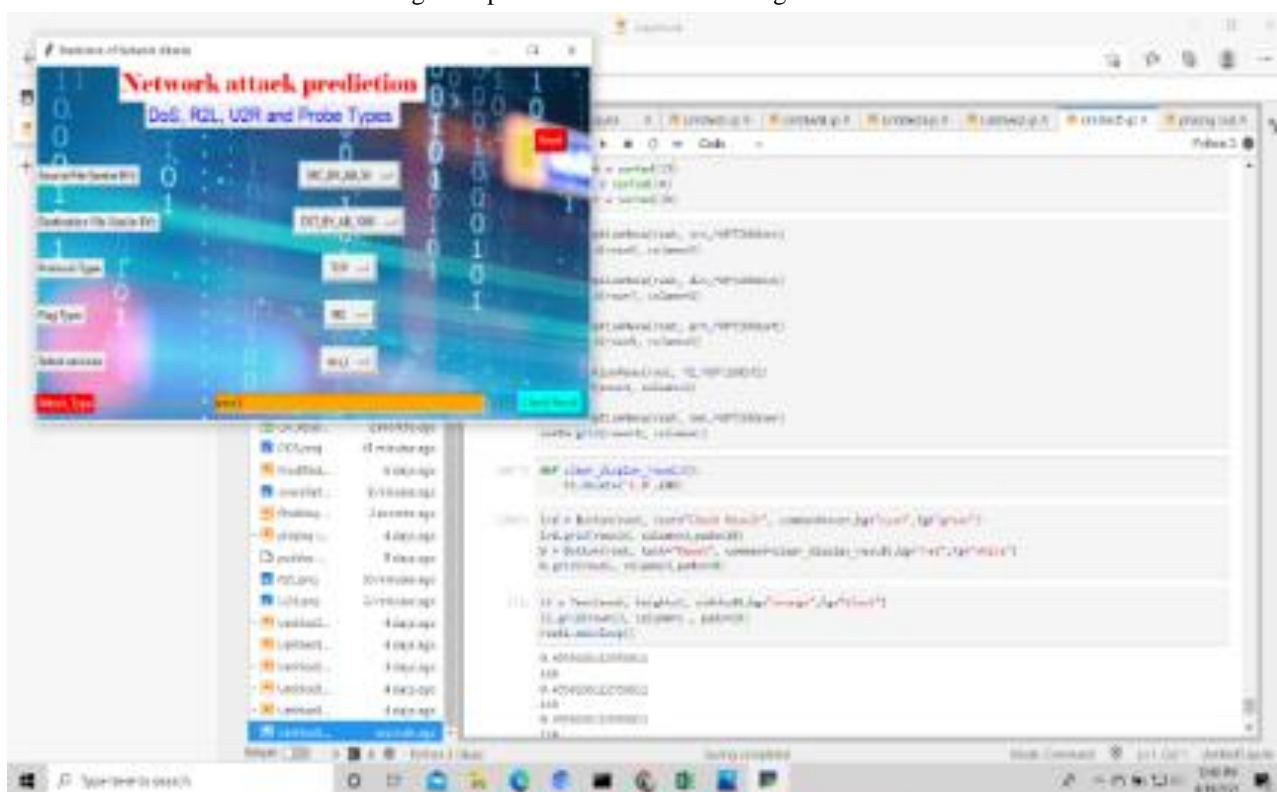


FIG: The Interface of Network Attack Prediction

The graphical User Interface of the network attack predictor is shown above where the attacks are predicted based on the user selection of the options given. The properties given are source file size in bytes, destination file size in bytes, protocol type, flag type and service. The attack is predicted based on user options. The result is obtained once the options are selected and reset and then the check result button is clicked.

7. Conclusion

The investigation thus far is identifying attackers' packets that may be recorded in a log file with sufficiently high success rates and without compromising users' privacy.

Learned the following by analysing the detection results:

1. The protocol percentage of log file is obtained in which icmp protocol is found to be most prominent with 83.20%.
2. In log file all attacks are identified best by KNN, SVM and random forest techniques yielding 99.6%.

Future Scope

Analysis of log knowledge is difficult facet of log management. The effective thanks to gain are to review and analyse logs wherever analysis method may be machine-driven by filtering log entries.

Purpose for filtering is to confirm that the manual analysis performed by directors is prioritized. Another effective technique is to own 2 reports for many vital entries and poorly understood entries whereas prioritizing review of the prior one. The review of latter permits in increasing and redefining work.

References

- Amit Pratap Singh “Effective and Efficient Data Aggregation Technique for Common Web Log XML Format for Digital Forensic Investigation (DFI) *International Journal of Engineering Research & Technology (IJERT)*, April-2015.
- Arpit Shah, Nilesh Kakade” A Survey: Investigation for Apache Log Preprocessing in Web Usage Mining” *International Journal for innovative research in multidisciplinary field*” 5, May – 2017
- Atif Ahmad and Tobias Ruighaver “*Design of a Network-Access Audit Log for Security Monitoring and Forensic Investigation*”, 2003.
- Bahalul Haque, Sharaban Tahura Nisa, Md. Amdadul Bari, Ayvee Nusreen Anika “Anonymity Network Tor and Performance Analysis of ‘ARANEA’ – an IOT Based Privacy-Preserving Router” *AKM 2019*
- Brian Cusack and Raymond Lutui “*Including Network Routers in Forensic Investigation*”, 2013.
- Chodhary Ravi Singh “Analysis of Router Poisoning using network attacks *International Research Journal of Engineering and Technology (IRJET)* Oct 2018.
- C. Reis, G. Parca, M. Bougioukos, A. Maziotis, S. Pinna, G. Giannoulis, H. Brahmi, P. André, N. Calabretta, V. Vercesi, G. Berrettini, C. Kouloumentas, A. Bogoni, T. Chattopadhyay, D. Erasme, H. Avramopoulos, and A. Teixeira J. “Experimental Analysis of an All-Optical Packet Router” *Opt. Commun. Netw*, vol 6, July 2014
- Dimitrios Xanthidis and Professor Dr David Nicholas “An Investigation of the employment or not of web log analysis by online businesses” *International Conference E-Commerce* 2007.
- Emmanouil Karamanos “*Investigation of home router security*” Master of Science Thesis Stockholm, Sweden 2010.
- Emmanuel Pilli, Ramesh Joshi and Rajdeep Niyogi “Router and Interface Marking for Network Forensics” *Advances in digital forensics* 27 Jul 2017.
- E. Aharoni, S. Fine, Y. Goldschmidt, O. Lavi, O. Margalit, M. Rosen-Zvi and L. Shpigelman “Smarter log analysis” *Digital Object Identifier*, 2011.

Friday Yakubu, Abdullahi Mohammed, Ibrahim Abdullahi “Correlation Analysis of Data Rate and Router Buffer Size on TCP Performance using OPNET Simulator” *International Journal of Computer Applications* September 2011.

George asquith, Charles Gibson “Basic well log analysis”, *The American association of petroleum geologists* 2012.

Giorgio Maria Di Nunzio, Johannes Leveling and Thomas Mandl “Multilingual Log Analysis”. *CLEF* 2010.

Howard F. Lipson “Tracking and Tracing Cyber-Attacks: Technical Challenges and Global Policy Issues” *Networked Systems Survivability Program*, November 2002.

Jan Valdman. “*Log File Analysis*” Technical Report No. DCSE/TR20004, July 2001.

Jie Li and Shi Zhou “*Performance Analysis of Multipath BGP*” March 16, 2021.

Karen kent, murugaih suppiyya. “*Guide to Computer Security Log Management*”. National Institute of Standards and Technology.

Kevin Vermeulen and Stephen D. Strowes “Multilevel MDA-Lite Paris Traceroute” *ACM Internet Measurement Conference*, 26 Sep 2018.

M. AlSabbagh, Jianping Chen, Guiling Wu and Xinwan L “Time Delay Analysis in Wavelength Router Optical Burst Switching (WR-OBS) Networks Haider”. *Journal of Optical Communications* 26, 2005.