Applying Big Data Analytics in DDos Forensics: Challenges and Opportunities

Augusto Gonzaga Sarmento, Kheng Cher Yeo, Sami Azam, Asif Karim, Abdullah Al Mamun, and Bharanidharan Shanmugam

Abstract DDoS (Distributed Denial-of-Service) attacks greatly affect the internet users, but mostly it's a catastrophe for the organization in terms of business productivity and financial cost. During the DDoS attack, the network log file rapidly increases and using forensics traditional framework make it almost impossible for DDoS forensics investigation to succeed. This paper mainly focuses on finding the most suitable techniques, tools, and frameworks in big data analytics that help forensics investigation to successfully identify DDoS attacks. This paper reviewed numbers of previous research that related to the topic to find and understand general terms, challenges and opportunities of using big data in forensics investigation. The data mining tools used in this paper for simulation was RapidMiner because of its ability to prepare the data before the analysis and optimizes it for quicker subsequent processing, and the dataset used was taken from University of New Brunswick's website. Algorithms that were used to evaluate the DDoS attack training dataset are Naïve Bayes, Decision Tree, Gradient Boost and Random Forest. The evaluation results projected that the majority of algorithms has above 90% of accuracy, precision and recall respectively. Using the data mining tools and recommended

A. G. Sarmento · K. C. Yeo · S. Azam · A. Karim (✉) · B. Shanmugam
College of Engineering, IT and Environment, Charles Darwin University, Casuarina 0810, NT, Australia
e-mail: asif.karim@cdu.edu.au

A. G. Sarmento
e-mail: amgsarmento@yahoo.com

K. C. Yeo
e-mail: charles.yeo@cdu.edu.au

S. Azam
e-mail: sami.azam@cdu.edu.au

B. Shanmugam
e-mail: bharanidharan.shanmugam@cdu.edu.au

A. Al Mamun
Institute of Information Technology, Jahangirnagar University, Dhaka, Bangladesh
e-mail: abdullah.iiuceee@gmail.com

H. Jahankhani et al. (eds.), *Cybersecurity, Privacy and Freedom Protection in the Connected World*, Advanced Sciences and Technologies for Security Applications, https://doi.org/10.1007/978-3-030-68534-8_15

algorithms will help reduce processing time associated with data analysis, reduce cost and improve the quality of information. Future research is recommended to install in an actual network environment for different DDoS detection models and compare the efficiency and accuracy in real attacks.

Keywords DDoS attacks · DDoS forensics · Big data analytics · Bid data forensics · Forensic investigation

1 Introduction

All information about network, protocols, application, and web are stored in a log file and this log file usually saves indiscriminately everything [1]. As we all aware that the network traffic is continuously increasing, which means that the size of logs files also increasing. Since all the information regarding DDoS attacks also stores in log files, investigation to find some meaningful insight regarding the attackers' details has become extremely difficult due to the big amount of data. Furthermore, using the current conventional forensic investigation method is time-consuming, costly and sometimes impossible to succeed.

Distributed Denial-of-Services (DDoS) is a type of cyber-attacks to an organization network where multiple systems flood the resources or bandwidth of the organization's systems. Malicious people use multiples zombie's computers to overwhelm the network's available resources which could be application or service with the request so that legitimate users not able to access the system [2]. This greatly affects the internet users in a computer network but mostly it's a catastrophe for the organization in terms of business productivity and financial cost. This is an ongoing issue for government agencies, financial institutions or any organization that need to be prevented and solved with a watchful approach [3].

This paper aims to identify the most suitable techniques, tools and framework in big data analytics that help forensics investigation to successfully identify DDoS attacks. This paper will deliver suitable data mining tools that will facilitate the forensic investigation in DDoS attack using big data, good forensics investigation methods that will be more suitable for big data investigation and also a report of experiment's result.

The research is focusing more on three things such as DDoS attacks itself (why and how it happens and also how to prevent it), big data analytics in forensic investigation and investigates DDoS attacks using big data analytics in forensics investigation. In addition, different data mining tools were evaluated and the chosen one was used in this paper. Moreover, current DDoS forensics methods were explored, and also numbers of algorithms that are used in DDoS forensics was explored as well. It is assumed that the evaluated data mining and algorithms will help to reduce cost and time spending on forensics investigation as well get an insightful pattern.

The rest of the paper is organized as follows: Sect. 2 provides the literature review of background knowledge regarding DDoS attacks, big data and forensics investigations. Section 3 discusses the methodology used in this paper. Section 4 presents the result of dataset evaluation using data mining tools and algorithms. Section 5 talks about the recommendation and discussion. Finally, Sect. 6 concluded the paper.

2 Literature Review

Several researches that have been focused on how difficult and challenging it is to do forensics investigation on big data. There are a number of solutions that have been proposed as well to overcome those difficulties and challenges and those solutions will be discussed more details in the later sections of the paper. In this section we elaborated on some related works that have been done prior to this paper alongside their statement and explanation. The evaluation outcome of related research attempts is arranged in 6 sections: information about digital forensics and its framework in Sect. 2.1; big data and its characteristics in Sect. 2.2; big data forensics and its challenge in Sect. 2.3; DDoS attacks and DDoS forensics in Sect. 2.4; algorithm used in related work in Sect. 2.5; and data mining tools comparison in Sect. 4.1.

2.1 Digital Forensic

Digital forensics is part of forensic science that responsible to identify an incident along with collection, examination, and analysis of evidence data. It is also responsible for investigating the cyber-crime and cyber-incidents, find the possible evidence and present it to the court for further judgment. Digital forensic has four main frameworks process [4–7]:

Identification

In this step, the investigator identifies the evidence of the crime or incidents and prosecute litigation. This step usually considered as the stage of preparation and preservation as well. The preparation includes preparing the tool, resources alongside with the necessary authorization or approval to collect data. Preservation involves securing the crime or incidents and possible evidence.

Collection

In this step, the investigator team starting to collect physical and digital evidence at the crime scene. Everything will be recorded in this stage and all the evidence is collected using standardised techniques. In this stage, while collecting the data, the investigator needs to make sure preserving the confidentiality and integrity of the data.

Organization

In this step, the investigation team efficiently collects the evidence which can lead to finding information regarding the criminal incidents. First, the investigators examine the collected data to find the potential pattern that can lead to the crime and the suspect. After that, the investigators analyse the correlation between found patterns and suspect to determine the fact.

Presentation

The investigator prepares the report of the result to present it in the court to prosecute litigation. The investigators have to make sure that the result they present must be easy to understand without requiring any specific knowledge.

2.2 Big Data

Nowadays, people define big data as a dataset that is too big, too fast and too difficult for traditional tools and frameworks to process. Big data is characterised by followings [4, 5]:

Variety

It describes different data that exist. Since big data comes from multiple sources like network or process logs, web pages, social media, emails, and any other various sensors, the data can be categorised as structured, semi-structured and unstructured.

Volume

It refers to the large amount of data that can be generated and stored. For example, in this era, many organizations like Google and Woolworths deal with terabytes or petabytes of data.

Velocity

Velocity refers to how big data getting bigger due to the new different systems that come every day. The velocity can be categorised as a real-time, batch, stream, etc. It is not only referring to the speed of incoming data but also about the speed of data flow inside the system.

Veracity

Veracity refers to the integrity and confidentiality of the data. It also involves data governance, quality of data and metadata management alongside the legal concerns.

Value

It refers to how big data can be turned to something that valuable for economy and investigation. Bid data can reveal all the important pattern that is searched for which is previously unknown and those can lead to something that valuable.

2.3 Big Data Forensic and Its Challenge

Big data forensics is defined as a branch of digital forensics that deals with evidence identification, collection, organization, and presentation to establish the fact using a very large-scale of dataset. Big data forensics can be looked at from two perspectives: first, a shred of small evidence can be found in the big dataset and second, by analysing big data, a crucial piece of information can be revealed [4–7].

To enable high-velocity capture, discovery, and/or analysis and to efficiently extract patterns and value from large volume and a wide variety of data, big data requires a new design generation of technology and architectures. Unfortunately, digital forensics' traditional tools and technologies are incapable of handling big data. Following are the challenges that encounter is each step of digital forensics investigation when dealing with big data [4–7]:

- *Identification*: When the amount of possible evidence is very large, it can be difficult to identify the important pieces of evidence to determine the fact.
- *Collection*: If there is an error that occurs during the collection stage, it will affect the whole investigation process. Because the *Collection* is considered as the most crucial steps.
- *Organization*: Since the existing analysis techniques do not comply with the characteristic of the big data, it can be challenging to organize big data set and identify the facts about the incidents.
- *Presentation*: It will be hard for the jury to understand the technicalities behind filtering, analysing big data and identifying value. Because it is not as easy as traditional computer forensics.

2.4 DDoS Attacks and DDoS Forensic Methods

The following details are taken previous related works that were conducted by [8–10]. The authors explained about DDoS attacks and DDoS forensic very precise and understandable. Table 1 summarizes the DDoS attack architectures from previous related works, Table 2 summarizes launching steps. There is not a lot that can be done apart from disconnecting the victim system from the network and fix it manually when DDoS attacks occur. However, the defence mechanism can be used to detect the DDoS as soon as possible and prevent it immediately, showed in Table 3. Table 4 summarizes DDoS Detection strategies and Fig. 1 showed the classification of DDoS defense mechanism. Also, Table 5 summarizes different algorithms used in the works reviewed earlier.

Table 1 Summary of DDoS attack architectures from previous related work

Attacks architecture	Description
Agent-Handler architecture	It also considers as Botnet based architecture. The attackers use the Botnet to conduct an attack and the Botnet consist of masters, handlers and bots
IRC (Internet Relay Chat)-based architecture	Instead of doing an attack using the original address, the attack is launched through a public chat system. Because IRC allows users to communicate without requiring any authentication check and no security
Web-based architecture	The attackers launch the attack by hidden themselves within legitimate HTTP and HTTPS traffic

Table 2 Summary of DDoS attack launching steps from previous related work

Steps	Description
Discover vulnerable host and agents	Attackers using tools and resources to find any system of the network that does not run with the antivirus virus and weak security defence system
Compromise	After the attackers finding the vulnerable system, they exploit the vulnerable system and install the attack code
Communication	The attackers communicate with the agents to schedule attacks, to identify active agents or to upgrade agents. The communication can be done via TCP, UDP and ICMP
Launching an attack	The attackers select the victim system and launch the attack

2.5 Algorithm Used in Related Work

See Table 5.

3 Methodology

To measure the accuracy and compare the efficiency of different learning models in detecting the DDoS attack, this paper utilised simulation. The simulation has been used in the past researcher that related to the same topic as this paper. What the past researches did are calculating the percentage of true negative, true positive, false negative and false positive. The dataset was divided into two parts throughout the simulation such as training as testing. Using this approach will help to simplify the complexity of DDoS attacks. Instead of capturing the real net flow data of DDoS attack, simulation aids in simplifying the data gathering process. Off course that this

Table 3 Summary of DDoS defence architectures from previous related work

Defence architectures	Description
Source-end defence mechanism	To prevent network users from generating the DDoS attacks, the source-end defence mechanism is deployed at the source of the attack. In this approach, all the malicious packet is identified by a source device in outgoing traffic and filter the traffic
Victim-end defence mechanism	It filters, detects or rate malicious incoming traffic at the routers of victim networks for instance network providing Web services. In this detection system, an anomaly intrusion detection system can be used
Core-end or intermediate router defence mechanism	Any router in the network can try independently to identify the malicious traffic and filter the traffic. For example, it is a better place to filter the traffic because both attack and legitimate packets arrive at the router
Distributed end or hybrid defence mechanism	One of the best strategies against DDoS attacks could be attack detection and mitigation at the distributed end. The core-end is suitable to filter all kinds of traffic and the victim-end can detect traffic accurately

Table 4 Summary of DDoS detection strategies from previous related work

Strategies	Description
Statistical	Utilizing the statistical properties of normal attack patterns for DDoS attacks' detections. Calculate a general statistical model for normal traffic and used it to test the incoming traffic to determine if it is legitimate traffic or not
Soft computing based	Using learning paradigms such as ANN (Artificial Neural Networks) which has self-learning characteristics to identify unknown disturbance or attacks in a system
Knowledge based	The rules that already established in advance are used to test against network events or actions. All the known attacks are defined as attack signatures and use the signatures to identify the actual attack
Data mining and machine learning	Protecting network devices and applications using an effective defensive system called NetShield from becoming a victim of DDoS flood attacks. It eliminates vulnerabilities of the system on the target machine using preventive and filter and protecting IP-based public networks on the Internet

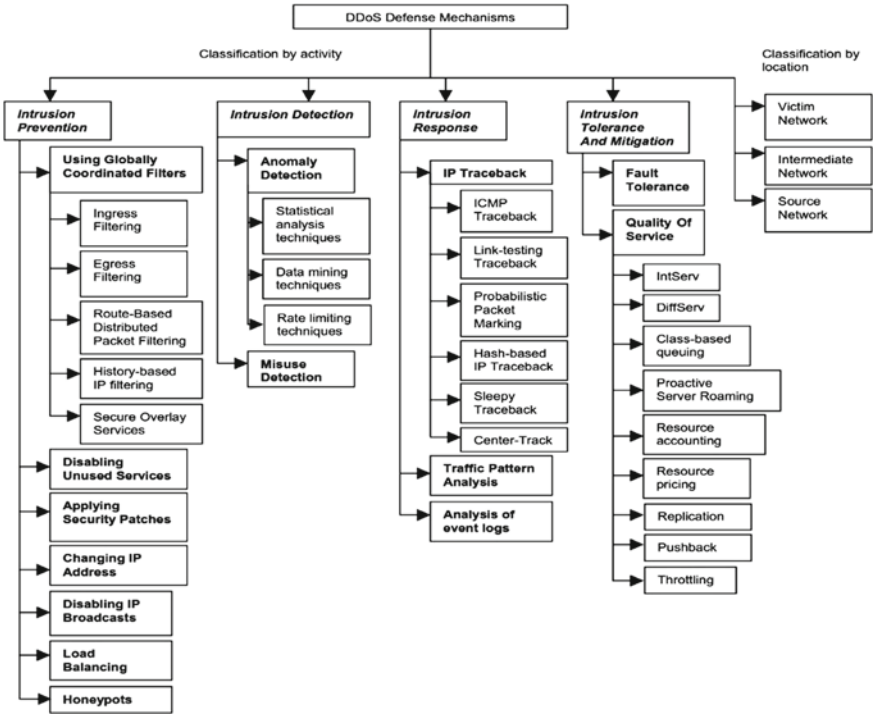

Fig. 1 Classification of DDoS defense mechanism

Table 5 Different types of algorithm using in previous related work

Algorithms	Paper and Authors
MapReduce by Hadoop	"DOFUR: DDoS Forensics Using MapReduce" by [1]
Hadoop Distributed File System (HDFS)	"Digital Forensics in the Age of Big Data: Challenges, Approaches, and Opportunities" by [4]
MapReduce, Decision Tree and Random Forest, Image Forensics, Neural Network and Neural Language Processing (NLP)	"Digital Forensics as a Big Data Challenge" by [7]
Decision Tree, Baysian, Neural Network, Nearest Neighbour, Genetic Algorithms, Case-based Reasoning, Rough Set and Fuzzy Logic	"Dealing with Terabyte Data Sets in Digital Investigation" by [11]
Gaussian Naïve Bayes	"A Novel DDoS Attack Detection Based on Gaussian Naïve Bayes" by [2]

approach comes with its drawback which is it may over-simplify the real situation of DDoS attack. Because the dataset that was used in this paper have been pre-processed before training and testing whereas in actual situation it is not. In addition, the constant changing actual threats environment may not reflect in the captured dataset.

The dataset that is used in this paper is taken from the University of New Brunswick (Canadian Institute for Cybersecurity)'s website. The dataset is divided into 7 different groups according to a different type of DDoS attack respectively as follows: Portman, UDPLags, LDAP, NetBIOS, UDP, MSSQL and Syn. The variables that is used to determine the DDoS attacks are time stamp, source and destination IP address. Since the datasets have been pre-processed and labeled, the data is ready to be evaluated using data learning algorithm and data mining tools.

RapidMiner is used in this study and it has been used extensively in data science. It is best in the area of future predictive analytics because it predicts future development based on collected data. The program can import Excel tables, SPSS files and data sets from many databases. In addition, it can be used for data mining, text mining, opinion mining and sentiment mining.

Power BI was used to confirm or visualize whether the dataset contains DDoS attack packet. The dataset was grouped by timestamp and then count the number of packets per timestamp. The DDoS attacks can be verified as shown on the spikes or the sudden increase in the number of packets (see Sect. 4 Part II). Power BI was chosen because it is more intuitive than the RapidMiner built-in visualization tools.

4 Results

The results are arranged in 3 sections, different data mining tools are compared in Sect. 4.1; finding the characteristics of DDoS attacks using Power BI in Sect. 4.2; and the result of dataset evaluation to find the accuracy, precision and recall of the algorithms in Sect. 4.3.

4.1 Software Comparison

Before deciding which data mining tools to be used for the evaluation, different data mining tools have been examined and explored such as RapidMiner, WEKA, Orange, KNIME and SAS. The characteristics and support of the data mining tools are summarised in Fig. 2. All these data mining tools have libraries that can be extended and used in the programming language.

After evaluating the performance of different data mining tools, RapidMiner was chosen for Analysis. RapidMiner can design modular operator concept even for very complex problems. To describe the operator modelling knowledge discovery (KD) processes, RapidMiner uses XML. It can also take input and output for and from

Tools	Characteristics	Programming Language	Operating System	Price/License
RapidMiner	• It predicts future developments based on collected data. • The program can import. Excel tables, SPSS files, and data sets from many databases. • It prepares the data before analysis and optimizes it for quicker subsequent processing.	Java	Windows, Mac, Linux	Free but also cost based on Versions
WEKA	• It has many classification methods such as artificial neural networks, ID3, decision trees and C4.5 algorithms. • Its machine learning capabilities support major data mining task like association, classification, clustering and regression • It is really useful for teaching and research purposes.	Java	Windows, Mac, Linux	Free Software
Orange	• Without extension of prior knowledge, it creates appealing and interesting data visualizations. • Its machine learning support data mining task such as clustering, regression, classification and much more. • It has the capabilities of learning about user's preference over time and reacts accordingly.	C++ Python (Extensions and query language)	Windows, Mac, Linux	Free Software
KNIME	• Helps to reveal hidden data structures. • Enables data mining and numbers of machine learning's methods to be integrated. • It is really effective when pre-processing data for example: loading data and extracting transforming.	Java	Windows, Mac, Linux	Free Software
SAS	• One of the best data mining tools for business analytics. • Good for large presentation in terms of prognostic sector and interactive data visualization. • It has high scalability so it can possible increase the performance proportionally by adding additional harder or other resources.	SAS language	Windows, Mac, Linux	Limited freeware through educational institutions.

Fig. 2 Data mining tools evaluation

any different form of dataset. RapidMiner has more than 100 learning schemes for clustering task, classification and regression.

4.2 Timestamp Visualization of Dataset Using Power BI

The graphs below shows the number of requests per protocol for each second. Using Power BI, the visualization is achieved by creating timestamp bins one second in duration. Then a measure is calculated as the count of records based on the column named "Flow ID". For each graph, the total of BENIGN (not harmful or safe) packets are shown to visualize what normal series of packet looks like before or after DDoS attacks. For the dataset used, this attack occurred in 3rd November 2018.

Figure 3 shows the flow of Portmap packets suddenly increases at around 10:01 am reaching 8.2 thousand request per second.

Figure 4 shows the flow of UDPLags packets suddenly increase at around 11:29–11:31 am reaching 13.5 thousand request per second. There is also an increase in UDPLags although the increase on other protocol is more significant. Overall, algorithm is able to identify BENIGN from malicious packets with high accuracy, precision and recall.

Figure 5 shows the flow of LDAP packets suddenly increase at around 10:21–10:27 am reaching 19.7 thousand request per second.

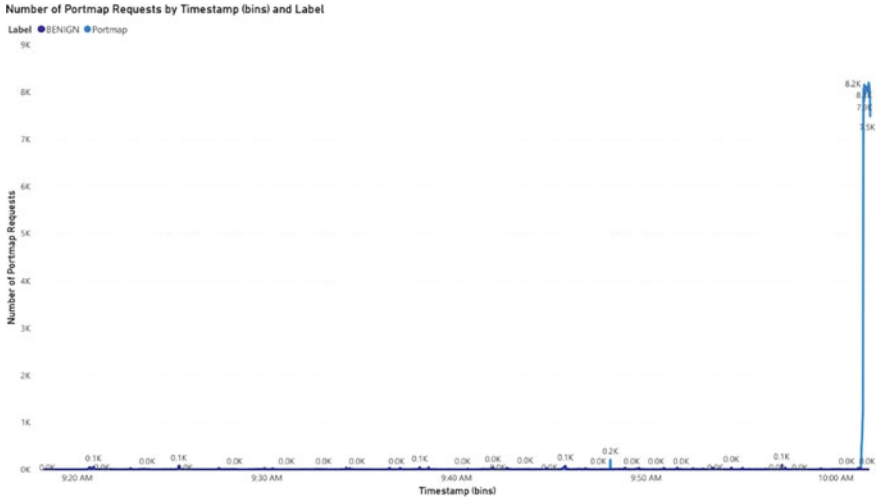

Fig. 3 Portmap DDoS attacks and timestamp

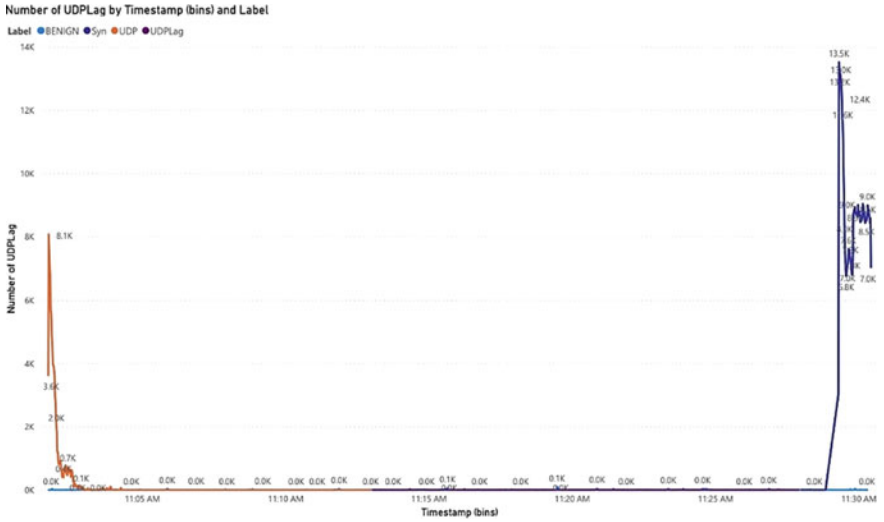

Fig. 4 UDPLags DDoS attacks timestamp

Additionally, Fig. 6 shows the flow of NetBIOS packets suddenly increase at around 10:02–10:09 am reaching 8.1 thousand requests per second.

Figure 7 shows the flow of UDP packets suddenly increase at around 10:53–11:01 am reaching between 9.3 and 11.7 thousand requests per second.

Figure 8 shows the flow of SYN packets suddenly increase at around 11:35–11:37 am reaching 13.6 to 27.2 thousand request per second.

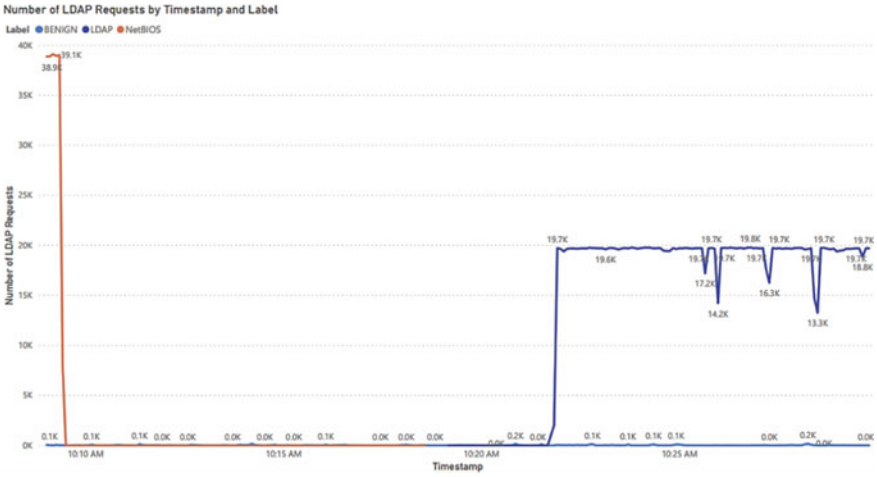

Fig. 5 LDAP DDoS attacks and timestamp

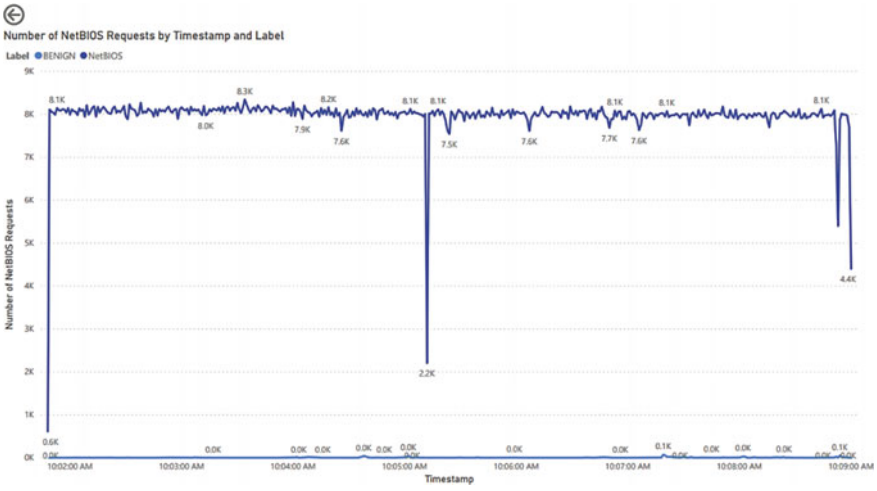

Fig. 6 NetBIOS DDoS attacks and timestamp

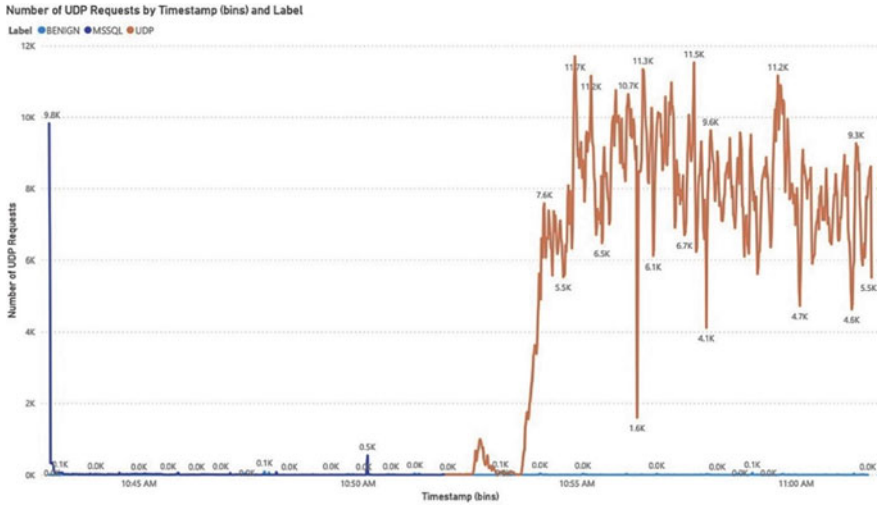

Fig. 7 UDP DDoS attacks and timestamp

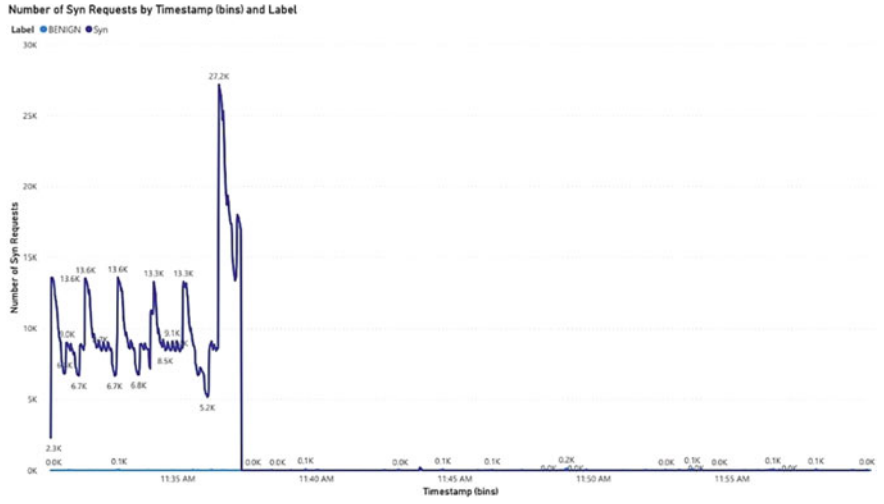

Fig. 8 SYN DDoS attacks and timestamp

Figure 9 demonstrates the flow of MSSQL packets suddenly increase at around 10:34–10:42 am reaching between 11.9 and 12.5 thousand requests per second.

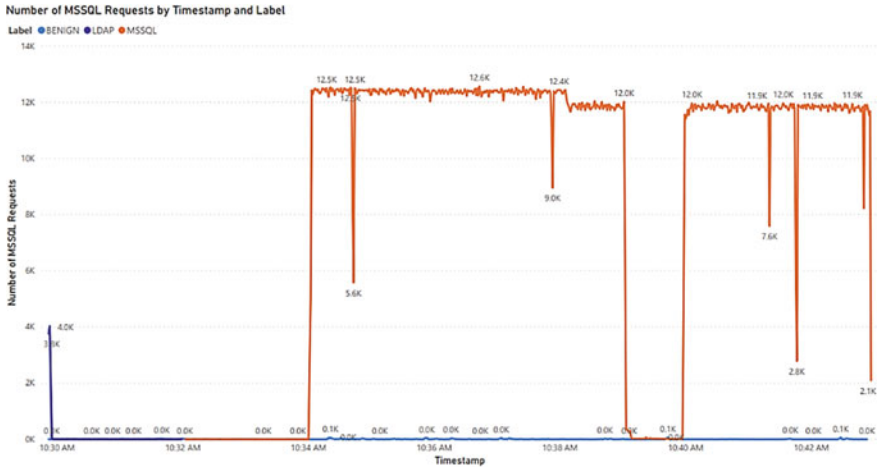

Fig. 9 MSSQL DDoS attacks and timestamp

4.3 Comparison of Machine Learning Algorithms

Distributed Random Forest.

Random Forests are based on "classification trees" which trains a 'forest' of decision trees and performs binomial classification predictions by introducing a training input to the individual trained trees in the 'forest' and promoting the dominant classification for each tree as the prediction result. In the distributed implementation of Random Forests, each cluster node is reassigned an identical division of the whole training dataset. Each computing cluster then trains an individual Random Forest cluster and majority classification for each cluster is identified as the prediction result [12].

Decision Tree.

Decision trees categorize the training data by sorting them from the root of the tree down to some leaf node, with the leaf node as the prediction result. Each leaf node in the tree serves as a test case for the highlighted attribute, and each path to the root is the possible answer to the test case. This algorithm is naturally recursive and is iterated for every subtree starting at the branch node. Decision trees use a variety algorithm to strategically decide where and how many splits to make. Each split increases the homogeneity of consequent splits. The integrity of the node increases depending on the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most consistent splits [13].

Gradient Boosting Machines.

Gradient Boosting trains many models in a steady increasing pattern. Gradient boosting performs by using gradients in the loss function $y = ax + b + e$ where e is

the error variable. The loss function is a measure indicating how good the model's coefficients are at fitting the underlying data. A logical understanding of loss function would depend on what we are trying to optimise. One of the biggest motivations of using gradient boosting is that it allows one to optimise a user- specified cost function, instead of a loss function [14].

Naive Bayes.

Naive Bayes (or Idiot Bayes) is a classification algorithm for binomial and polynomial classification problems. The calculation of the probabilities for each hypothesis is simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P (d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) * P(d2|H). The approach executes well on data where the assumption that the attributes do not interact is disregarded [13].

The precision is reliable but still depends on the split between training data and testing data. For instance, for MSSQL requires 50/50 split in order to get result for Decision Tree and Naïve Bayes and NetBIOS requires 90/10 split to get result for Gradient Boosting Machine and Naïve Bayes. The split is needed to change because for above dataset, there is not enough information to accurately train the model. For the rest of the dataset, only requires 30% training and 70% testing to get an accurate and precise result.

Since recall is close to 100%, as shown in Table 6, most of the true positive was found therefore proving that the training covers almost all the dataset. Since the accuracy is also almost close to 100% for the average, it simply means that the models predict most of the data correctly. In most test except for the two (Naïve Bayes' precision for NetBIOS and SYN dataset), precision close to 100% means how useful the generated model is. Generally, the algorithms that are used in this simulation paper can be used to examine DDoS attacks.

5 Discussion and Recommendation

After evaluating the different machine learning algorithms, for those algorithms that resulted in high precision, accuracy and recall, it can be recommended to use the algorithm model for DDoS forensics investigations. For lower values, further modelling is required to generate a model that is accurate and precise enough for it to be used in DDoS forensics investigation. It is also recommended for further research to use different dataset and up to date tools to confirm the findings of this research.

After evaluating previous research papers and related work, it can be said the traditional forensic framework is not suitable for big data investigation or DDoS forensics. Khattak et al. [1] and Zawoad and Hasab [4] proposed to use Hadoop's MapReduce for the forensic investigation of DDoS attacks. This method will help to find out whether the system is under attack, who attacks the system and which

Table 6 Algorithmic performance

System	Algorithm	Accuracy (%)	Precision (%)	Recall (%)
Portmap	Distributed Random Forest	99.97	100.00	99.97
	Decision Tree	99.73	99.76	100.00
	Gradient Boosting Machine	96.30	100.00	95.73
	Naïve Bayes	99.91	99.97	97.90
UDPLags	Distributed Random Forest	99.93	100.00	99.93
	Decision Tree	99.93	99.97	94.10
	Gradient Boosting Machine	100.00	99.92	99.43
	Naïve Bayes	99.93	100.00	99.93
LDAP	Distributed Random Forest	99.99	98.21	99.74
	Decision Tree	99.99	100.00	99.99
	Gradient Boosting Machine	100.00	99.93	98.31
	Naïve Bayes	99.92	100.00	99.92
NetBIOS	Distributed Random Forest	–	–	–
	Decision Tree	100.00	100.00	99.49
	Gradient Boosting Machine (90/10)	100.00	100.00	97.54
	Decision Tree	100.00	100.00	99.49
UDP	Distributed Random Forest	–	–	–
	Decision Tree	100.00	94.56	99.79
	Gradient Boosting Machine	99.92	100.00	100.00
	Naïve Bayes	99.96	100.00	99.96
SYN	Distributed Random Forest	–	–	–
	Decision Tree	–	–	–
	Gradient Boosting Machine	100.00	100.00	99.40
	Naïve Bayes	99.36	56.37	100.00
MSSQL (50/50)	Distributed Random Forest	–	–	–
	Decision Tree	100.00	100.00	100.00
	Gradient Boosting Machine	–	–	–
	Naïve Bayes	99.94	100.00	99.94

incoming traffic is part of the attack. Hadoop provides MapReduce to use for parallel processing of distributed data. Adedayo [5] reassessed the digital forensic examination stages and proposed additional techniques and algorithms that help to handle big data issues in the investigation Fig. 10. The author continues stating that the proposed solution is not intended to stand alone rather than to support the existing framework and to solve the challenge facing by existing methods.

Another study conducted [2] talks about the DDoS attacks and the impacts. The authors proposed a new approach based on network traffic to analyse and detect

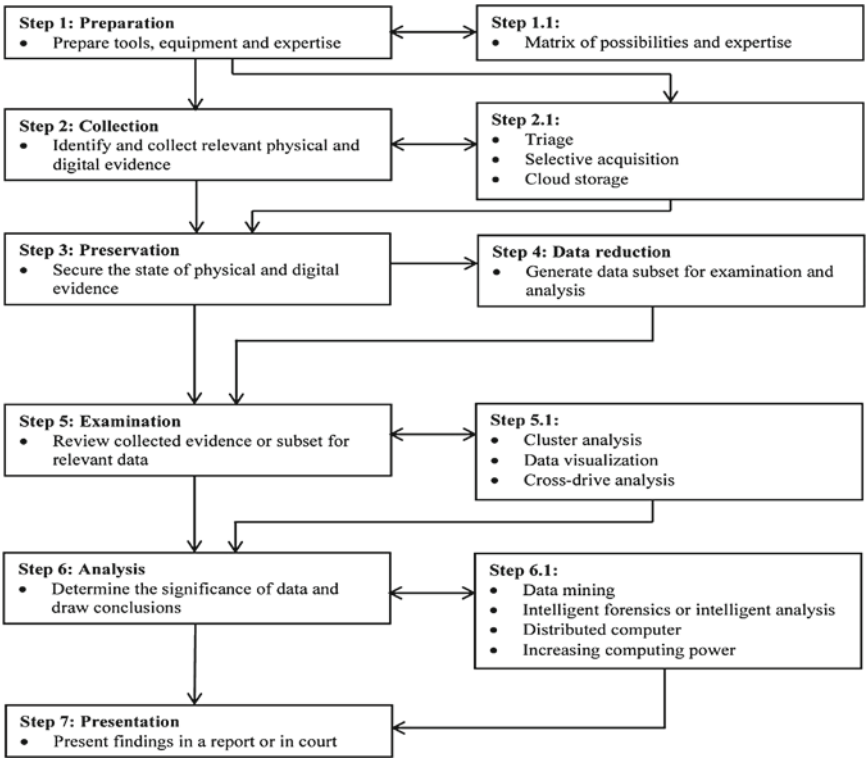

Fig. 10 Proposed digital forensics framework

DDoS attacks using Gaussian Naïve Bayes method. Hoon et al. [15] stated supervised learning algorithms such as Deep Learning, Gradient Boosting, Distributed Random Forest and Naïve Bayes performed better than unsupervised learning algorithms such as farthest first, canopy, make a density-based cluster and filtered cluster. To tackle the big data challenge, Guarino [7] suggested using decision trees and random forest to find anomalous behaviour or anomalous transaction and neural network to recognise application normal behaviours (it is suitable for network forensics to recognise complex patter). Beebe and Clark [11] said implementing data mining tools and research to the forensic investigation will help to reduce processing time associated with data analysis, reduce cost and improve the quality of information.

6 Conclusions

This paper proposes a more practical solution and framework to facilitate DDoS forensics investigation as illustrated in Fig. 10. This paper also carried out simulation using RapidMiner and compare different accuracy, precision and recall of the algorithms in detecting DDoS attacks. This paper evaluates 4 different machines learning

algorithm and compares its usefulness and effectiveness. This research initiative used RapidMiner, unlike the previous studies which is majority used WEKA because RapidMiner accept any data format and it prepares the data prior to analysis and optimizes it for faster subsequent.

References

1. Khattak R, Bano S, Hussain S, Anwar Z (2011) DOFUR: DDoS using mapReduce. Front Inf Technol 117–120
2. Fadil A, Riadi I, Aji S (2017) A novel DDoS attack detection based on Gaussian Naïve Bayes. Bull Electr Eng Inform 6(2):140–148
3. Kupreev O, Badovskaya E, Gutnikov A (2019) DDoS attacks in Q3 2019. SECURELIST [post-print]. https://securelist.com/ddos-report-q3-2019/94958/
4. Zawoad, S, Hasan R (2015) Digital forensics in the age of big data: challenges, approaches, and opportunities. In: 17th international conference on high performance computing and communication (HPCC). 7th international symposium on cyberspace safety and security (CSS), pp 1320–1325
5. Adedayo OM (2016) Big data and digital forensics. In: International conference on cybercrime and computer forensics (ICCCF), pp 1–7
6. Guo H, Jin B, Shang T (2012) Forensic investigations in cloud environment. In: International conference on computer science and information processing (CSIP)
7. Guarino A (2013) Digital forensics as a big data challenge. ISSE
8. Peng T, Leckie C, Ramamohanarao K (2007) Survey of network-based defense mechanism countering the DoS and DDoS problems. ACM Comput Surv 39(1):3-es
9. Prasad KM, Redy ARM, Rao KV (2014) 'DoS and DDoS attacks: defence, detection and traceback mechanisms—a survey. Glob J Comput Sci Technol: E Netw Web Secur 14
10. Douligeris C, Mitrokotsa A (2004) DDoS attacks and defense mechanism: classification and state-of-the-art. Comput Netw 44(5):643–666
11. Beebe N, Clark J (2005) Dealing with terabyte data sets in digital investigations. In: IFIP international conference on digital forensics, vol 194, pp 3–16
12. Breiman L (2001) Random forest. https://libguides.ioe.ac.uk/c.php?g=482485&p=3299839
13. Karim A, Azam S, Shanmugam B, Kannoorpatti K, Alazab M (2019) A comprehensive survey for intelligent spam email detection. IEEE Access 7:168261–168295
14. Towards Data Science (2018) Understanding gradient boosting machines. https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab
15. Hoon KS, Yeo KC, Azam S, Shanmugam B, Boer FD (2018) Critical review of machine learning approaches to apply big data analytics in DDoS forensics. In: International conference on computer communication and informatics (ICCCI-2018)