NET-DPI: Network Filter Using Deep Packet Inspection and Machine Learning Classification

Mareine Nassef
Faculty of Computer Science
Misr International University
Cairo, Egypt
mareine1412013@miuegypt.edu.eg

Ahmed El-Rawy
Faculty of Computer Science
Misr International University
Cairo, Egypt
ahmed1409549@miuegypt.edu.eg

Eng. Silvia Soliman
Faculty of Computer Science
Misr International University
Cairo, Egypt
silvia.wahballa@miuegypt.edu.eg

Mariam Mohie
Faculty of Computer Science
Misr International University
Cairo, Egypt
mariam1411450@miuegypt.edu.eg

Marwan Atef
Faculty of Computer Science
Misr International University
Cairo, Egypt
marwan1404082@miuegypt.edu.eg

Dr. Ghada Khoriba
Faculty of Computers and Information
Helwan University
Cairo, Egypt
ghada_khoriba@fci.helwan.edu.eg

***Abstract*—NET-DPI is an internet filter for an organization's network. It filters the organization's network traffic to achieve network resource optimization and a controlled environment set by its administrators. Most internet filters block entire websites if some or all of the website's content is deemed harmful/irrelevant; however, that means also blocking potentially beneficial content as well by default. The contribution discussed in this paper, is the achievement of accuracy and resource utilization by combining multiple techniques from different domains to solve this problem. This is done by reaching into more depths to access each web page and judging it on its own merits, instead of only taking action according to the website as a whole, or depending on the website's description. This paper explains the main steps to achieve these objectives: intercepting the network flow using firewall (with SquidGuard); extracting packets using Deep Packet Inspection (with tcpdump); then analyzing the content of the web page using a combination of machine learning and deep learning classifiers(K-Nearest Neighbor, Gradient Boosting, and Recurring Neural Network); and accordingly, reaching a decision to allow or block that web page.**

***Index Terms*—Networks, Deep Packet Inspection, tcpdump, Content Filter, Firewall, SquidGuard, Deep Learning, Classification, Recurrent Neural Network, K-Nearest Neighbor, Gradient Boosting**

I. Introduction

NET-DPI is a network filter; it is placed on an organization's network server. A network server acts as a Man-in-the-Middle between the network's clients in the organization and the website servers that offer internet services. The way the internet passes information from client to server and vice versa, is by using packets. A packet is a small partition of the data sectioned according to internet standards and protocols. Among one of the main techniques that NET-DPI uses is Deep Packet Inspection (DPI). DPI [6] is the extraction of the payload part of a packet, which contains the data. Whereas, Standard Packet Inspection is just the reading of packet headers, which only contain information about the data. DPI is considered as one of the most important parts in content-aware network applications [9]. After the data is inspected and extracted, it needs to be classified. Classification [14] is part of Machine Learning and it is giving a label to unlabeled input, according to previous training. There is also a more advanced set of techniques for classification called Deep Learning. In this paper, all of these techniques in different domains are used together to solve one problem.

The problem that NET-DPI solves is: the blocking of important web-pages depending on the websites they belong to; whether the website is considered a shady or irrelevant website, or the website is judged by the host-given information about its content, but in either case, the judgment is not based on the content of each web-page itself. NET-DPI filters the content of an organization's internet traffic using a firewall, inspects the content using DPI, uses analysis techniques to extract the semantics from the content, and finally uses classification algorithms to reach a decision whether to allow or block the web-page.
This paper discusses the main objectives of NET-DPI which are: finding a way to filter network traffic for organizations such as universities in a way that is both beneficial for the university and its students. The university achieves efficient and strict internet filtering, and the students can have access to educational and important content.

In this paper, related work is considered, reviewed, and analyzed. The work done in this paper is relevant to modern networking problems, and according to others [8], key parts of our proposed methodology and techniques are used to even solve other problems in the domain such as the enhancement of DPI. Also there are other researchers [16] that have a similar approach to the one in this paper; however, different algorithms

and data sets were used as well as different results were reached. A more intensive review of related work can be found in section II. In Section III, the methodology of the proposed system and its techniques are discussed. Furthermore, section IV is a discussion of the tools that have been used, their usages, and their results. Finally section V, the conclusion of the paper, in which is a final explanation of the link that was conducted between the problem and its solution.

II. Related Work

A. Network Filters and Deep Packet Inspection

The techniques used in this paper, and its methodology of using these techniques together, are relevant and considerably new as seen in paper [16]. The paper mentions the usage of similar techniques to achieve different results. The paper discusses how much network traffic filtering and classification have come to light due to the quick growth of online websites and applications. In addition to how much classifying network traffic has become a challenge. For that reason, the authors of the paper have proposed a solution, called "Deep Packet": a solution that combines both classification techniques and DPI. Deep Packet classifies network traffic using deep learning. Their classification algorithm is Neural Network (NN), for example CNN with an average accuracy 95%, as well as stacked autoencoder NN (SAE) with its average accuracy 95 %.

Furthermore, another related work [25] is one that advocates how much network filtering and classification is of vital importance for network management and network security. The paper discusses the concept of the majority of unknown traffic is conducted by certain types of applications; it gives this unknown traffic a name: Elephant Traffic. The authors state that traffic sharing the same server IP and the same server port is generated by the same application, and belongs to the same service. Therefore their proposed solution is a novel method, where statistical features are used for cluster analysis, to classify this Elephant Traffic. In order to filter this traffic, nDPI is used, which is an open-source DPI library. It is used to filter through this unknown traffic (Elephant Traffic).

In addition to filtering network traffic using deep packet inspection, the corresponding paper [8] demonstrates how certain patterns, because of their reiteration, can cause DPI to slow down the process of filtering traffic. Furthermore, these repeated patterns can be scanned only once and then skipped if encountered again. These patterns can be found whether in text or in bytes. Similarly, in this paper, the patterns used are of text, obtained from the page source. However computation is not a concern, and that is discussed further along in this paper. Also in further sections of this paper it is discussed that only certain traffic is analyzed and not all of it, so that is how this paper surpasses the slowness of the combination of the techniques.

B. Machine learning and Deep Learning

As mentioned in paper [22], automated classification has witnessed a boost in interest. The rapid growth of machine and deep learning has taken on added importance in the last 10 years. The authors of the paper state that the reason for this growth is due to the increased availability of digital documents and the increasing need to organize them. The purpose of their paper is to discuss the main approaches to text categorization and the usage of machine learning in it. Furthermore, in their paper it is demonstrated in details the classifier constraints. Moreover, the paper discusses how Text Classification (TC) is an instance of Text Mining. Furthermore, TC has been applied in several projects whether in document filtering, automated metadata generation.

	Enterprise	Malware
Algorithm	Standard Enhanced	Standard Enhanced
LinReg	99.92% 99.28%	0.00% 58.65%
L2-LogReg	93.35% 98.36%	16.86% 76.13%
L1-LogReg	92.75% 98.97%	19.71% 75.08%
DecTree	97.55% 97.02%	40.98% 83.33%
RandForest	99.53% 99.99%	33.54% 76.79%
SVM	11.94% 99.78%	77.98% 72.62%
MLP	95.90% 99.545	20.61% 72.53%

TABLE I
USED ALGORITHMS AND THEIR ACCURACIES [4]

Worthy of mentioning is another related paper [4] that uses DPI as well as classification algorithms. In their paper, the authors used real network data as an input for six common classification algorithms and their accuracies as shown in table II-B. Furthermore, these algorithms were not combined together to generate hyper optimized algorithms, on the contrary, each one is tested individually.

TABLE II
USED ALGORITHMS AND THEIR ACCURACIES FOR FIRST DATA SET [12]

	Word Appearance	Word Frequency
C4.5 (J48 IN WEKA	83.0%	80.5%
Naive Bayes	95.0%	94.5%
PRISM	74.5%	63.0%
Support Vector Machine	95.5%	94.5%

TABLE III
USED ALGORITHMS AND THEIR ACCURACIES FOR SECOND DATA SET [12]

Algorithm	Accuracy	AUC
Naive Bayes Multinomial	92.9%	0.954
SVC	77.5%	0.992
Linear SVC	98.9%	0.993

Finally, the authors in this paper [12] advocate that the fact that the internet has become a fast developing environment that is used in every organization (e.g. educational institution,

governments).The authors state that the students have the ability to surf the internet for educational content, but unfortunately they still have the ability to download and surf noneducational content which consumes bandwidth of the network. The majority of the time proxy server maintain a blacklist of Uniform Resource Locators (URL), which are kept as a static list. The authors proposed a solution which is to use machine learning to predict whether the URL is considered educational or noneducational. As a matter of fact, their algorithms have been tested on two data sets. The first data set had these algorithms: Naive Bayes and two Support Vector Machine (SVM) classifiers, namely SVM with RBF kernel (SVC) as shown in table II with their accuracies. As for the second data set, the algorithms that were used were: SVM with linear kernel (Linear SVC), Naive Bayes Multinomial SVC as demonstrated in table III. The input of the previous classifiers was the text inside the body tag found in Hypertext Markup Language (HTML) page sources after removing tags and extracting English words.

III. Methodology

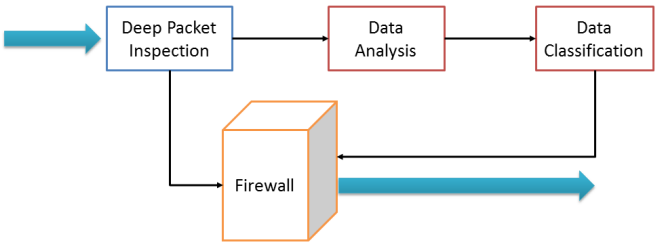

Fig. 1. NET-DPI Components

In figure 1, a wide overview is shown of the main components that make up NET-DPI's methodology and the flow between them, as well as the input and output of each phase. Each phase is explicitly explained in the following subsections, along with its techniques and purposes. As previously explained, NET-DPI is the combination of all of these components together. The way these components are used is the contributional aspect of this methodology. The network flow of the organization runs through the DPI filter, and if the desired traffic is detected (online video streaming website reply), the rest of the components start working. The firewall then stops the traffic from reaching its destination (a client using the network within the organization). Consequentially, the output of the DPI is then taken as input for the data analysis and Classification phases. The data is analyzed, then a decision is reached, and finally an action is taken by the firewall accordingly.

A. Deep Packet Inspection

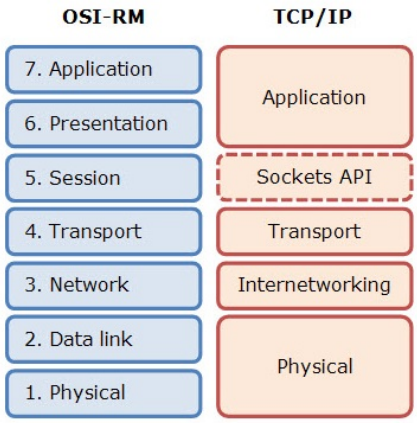

Fig. 2. The OSI and TCP/IP Layered Architectural Models of Modern Networking [1]

Packets travel across the network layer of the Transmission Control Protocol/Internet Protocol (TCP/IP) layer [5] (figure 2). As shown in Figure 3, Packets carry certain information that make it reach its destination; whether the sender's IP address, the intended receiver's IP address, something that tells the network how many packets has been broken into and the number of this particular packet as well as the data of the packets (payload). There are two ways to read these packets: Packet Inspection and Deep Packet Inspection. Packet Inspection [26] is the reading of the headers at the network layer, as shown in figure 2, to know information about the data (such as destination, source, size,..), without reaching the payload at all. Deep Packet Inspection (DPI) means reaching the packet through multiple layers, not just the network layer. Not only does this mean getting the payload, but also getting it in multiple forms: binary (from Physical Access Layer), ASCII(from Network and Transport Layers), or in the data's original form (Application Layer). DPI can be used for protocol detection, anti-virus, anti-malware and Intrusion Detection System (IDS) [19]. The purpose of DPI in this paper is protocol detection and data extraction. DPI is used to get the payloads of all packets of a certain internet transaction (one web page) and turning it to their original form: web page source (i.e. text). In conclusion, the DPI tool will inspect all the network flow of the organization and if any packets are found to be from the online streaming video website, then the firewall is called to pause these packets from reaching their destination until computations are made and a decision is taken.

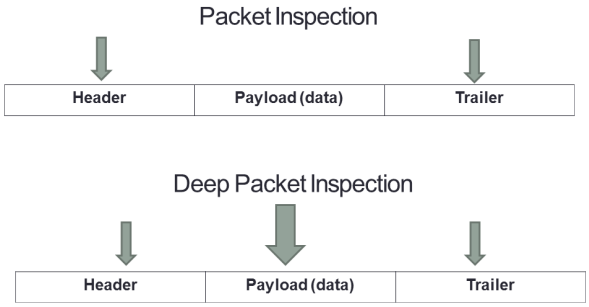

Fig. 3. The Difference between Standard and Deep Packet Inspection

B. Firewall

In computing, a firewall is a network security system that monitors and controls incoming and outgoing network traffic based on predetermined rules [7]. For NET-DPI, the rule is to allow any traffic except the rejected content categories provided by the organization administrators after applying it to the data analysis phases. For example, if the organization is a university and its Dean wanted only education videos to be allowed, then it is the firewall's task to block unwanted traffic of all other categories and allow the wanted educational traffic. However, the firewall cannot know the type of websites allowed or not; that is the data analysis phases' task. The type of firewall used for this system is called a "network firewall", also known sometimes as a "packet filter". Packet filters look at network addresses and ports of packets to determine if they must be allowed or blocked [20]. The firewall tool used for this system gets the traffic's URL from the DPI component, then pauses it from reaching the user whom had requested it. The time during the pause, the analysis phases are run over the page source outputted by the DPI component as well. When the analysis and classification are done, a decision is given to the firewall to allow or block that URL. The action of allowing or blocking is in fact redirecting to another URL. The firewall, resides as a part of the network server, and controls clients' traffic by redirecting web page request traffic to other kinds of traffic for example a block page.

C. Data

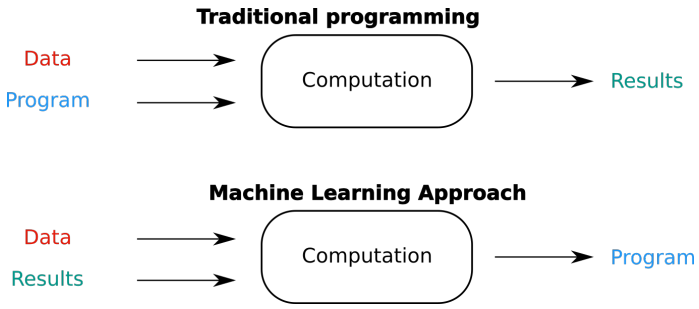

Fig. 4. The Difference between Traditional Programming and Machine Learning Approach [2]

The data contained in any video streaming page is either text or video, and for both, classification and analysis are used so that the system could "understand" the content. As shown in figure 4, Machine Learning is an approach used to find patterns and similarities and extract rules to follow. Classification is a part of the Machine Learning approach, and that part is called supervised learning. Classification is a training set of correctly identified observations given, so when the input is unidentified, the machine can still categorize it [22]. Deep Learning or Deep Neural is an advanced approach to Machine Learning, designed to act as an artificial Neural Network copying that of the human brain [17]. Both Machine Learning and Deep Learning classifiers use features extracted from the original content. However, Deep Learning classification algorithms can function without feature extraction, so they can work with data such as text, hexadecimal, or bytes, as well as encrypted content.

The main type of data that NET-DPI uses for web page content analysis is text. Text is usually unstructured, which means it cannot be the input to uniform computation. A web page source, which is the script that makes the web page appear as it does, is semi-structured because of its tags. However, for classification algorithms to use this text as input, it must be analyzed and structured first. That is called Feature Extraction. During the analysis phase, the text undergoes some Natural Language Procedures to convert it from unstructured to structured data, an example of that is shown in figure 5. These Procedures are: the removal of any parts of the text that is not part of the English language or any other language detected; then the removal of non-essential words that do not contribute to a sentence's meaning (such as: "the", "a", "an"), and stemming, which is returning all words to their respective origins ("entertainment", "entertaining", and "entertained" all become "entertain"); finally using a pre-prepared bag of words for each category to match the input and turn the unstructured text to structured numerical data. The categories below act as both features and classes.

- Film and Animation
- Autos and Vehicles
- Music
- Pets and Animals
- Sports
- Travel and Events
- Gaming
- People and Blogs
- Comedy
- Entertainment
- News and Politics
- How-to and Style
- Education
- Science and Technology
- Nonprofits and Activism

film and animation	education	autos and vehicles	entertainment	class
4	2	3	33	entertainment
17	10	0	55	entertainment
1	6	0	18	entertainment
0	7	0	22	entertainment
0	135	0	18	entertainment
0	0	0	19	entertainment
1	0	0	11	entertainment
2	0	0	16	entertainment
1	6	5	2	entertainment
0	4	0	74	entertainment
9	275	0	57	music
9	268	0	90	music

Fig. 5. An Example of the Initial and Final states of the Input Text

IV. Experiment

Different algorithms were tested and compared, as well as some ready-made libraries for technical efficiency. Bash Script for Linux Operating System is used to run all the commands and codes of the following technical experiments. Below in figure 6, is a simple representation of the input/output flow of NET-DPI from and to each phase.

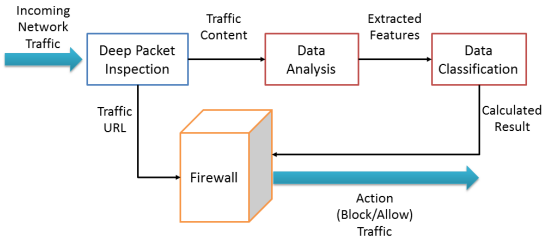

Fig. 6. Overview of NET-DPI with Input/Output Flow between Components

A. Deep Packet Inspection Phase

The tcpdump library is chosen for the DPI experiment; it is a library from Wireshark. The tcpdump library [23] is a command based library that monitors network traffic. The library allows the user to display TCP/IP and other packets being transmitted or received over a network as well as filtering traffic then outputs the payload as page source to be delivered to the next phases. Moreover, It is used before the firewall to detect the unwanted traffic and notify the firewall to block the URL extracted after getting the output from it (page source) and using it as an input for analysis and classification phase.

B. Firewall Phase

SquidGuard is considered a static web content filter; it depends on its blacklist to block or filter web content. Accordingly when it filters, it does not acknowledge the content of the web page, only its URL. In order to be able to dynamically block a web page, the firewall's blacklist should be frequently updated [12]. For NET-DPI, the blacklist will be dynamically handled and edited by saving the output of the other phases directly into its blacklist. SquidGuard is used to stop the traffic from reaching its end client until the analysis and classification are done, then it is given the command to block or allow according to their results. In fact SquidGuard's way to allow and/or block is through redirecting to the original URL (allow) or to a "page blocked" URL (block). Since SquidGuard does not have a "pause" command, it will redirect certain traffic to a "pause page" and then with outside code, the page will refresh then redirected according to the classification decision.

C. Classification Phase

The analysis and classification phases are probably the most important phase in the this section. In order to find the right classifier for the right job and get accurate results, there must be experimentation. Finding the right classification algorithm is an important component of many data mining projects [18]. Which requires careful thought of experimental design so as not to result in statistically invalid conclusions.

The experiments over the classification phase were done to find the best classifier for the problem at hand. The subject of the following experiments is the page source of an internet transaction. In order to prepare for the classification phase experiments, preprocessing was done i.e. the analysis phase. The first type of experiments in the classification phase is testing different classifiers over known data to see the accuracy for each. The following are the classifiers that were used separately.

- Machine Learning Classification Algorithms
  1) K-Nearest Neighbor (KNN): used for classification and regression with input that consists of the k closest training examples according to Euclidean Distance [3]
  2) Decision Tree:a decision support tool that uses a tree-like graph of decisions and their possible consequences. [21]
  3) Naive Bayes: one of the top 10 data mining algorithms due to its simplicity, efficiency, and efficacy. [24]

4) Support Vector Machine (SVM): achieves good performance when applied to real problems, especially text-categorization problems. [11]
5) Gradient Boosting: achieves state-of-the-art performance in academia, industry, and data analytics competitions. [13]

- Deep Learning, Neural Network Algorithms
  6) Recurrent Neural Network (RNN): a special type of neural network equipped with additional recurrent connections. [15]
  7) Convolutional Neural Network (CNN): a simple and fast algorithm, it introduces a new way to do unsupervised feature learning, and it provides discriminative features which generalize well. [10]

A small data set of approximately 4000 YouTube.com page sources of pre-categorized videos was testes for initial results. Then a much larger data set of approximately 148,000 of the same data was collected and tested. The following table shows the testing accuracy results for each classifier in both data sets. It is noteworthy that the data sets were divided into 80%-20%: 80% training and 20% testing (the 80-20 split is the most common one used for internal validation of the data set).

TABLE IV
CLASSIFICATION ACCURACY RESULTS

Classification Algorithm	**data set Size Variance**	
	with 4K	***with 148K***
KNN	100%	99%
Decision Tree	97 %	95%
Naive Bayes	22% *	85%
SVM	94%	92%
Gradient Boosting	99%	94%
RNN		92%
CNN	94%	93%

*This odd result is due to over-fitting since the categories were not evenly divided over the data set.

D. Experiment Conclusions

From the experiment done, the best classifying algorithms are: KNN, Gradient Boosting, and Decision Tree. It is best to use them together as a combination. In order for any input to be classified as "education", for example, then all three classifiers must classify it as such. The complexity of these algorithms is not a concern since all of these computations will run on a server.

V. CONCLUSION

In conclusion, even though there are applications that may be similar to NET-DPI whether the usage of DPI and/or classification algorithms, NET-DPI tackles these techniques with a different perspective as it reduces resource consumption as much as possible by interacting with application layer and/or Data-link/Network layers. Also, to be able to deliver smart access according to the content, without an absolute block of websites is a contribution of itself. As part of the future work for NET-DPI, multi-thread processing is considered so NET-DPI could handle multiple users at the same time. Furthermore, NET-DPI has quite a high potential to be dynamic, so that is NET-DPI's focus as future work. It can be applicable to any organization that has administrators that want to monitor and filter internet traffic over their inferiors in a more sophisticated way; it could filter any category from any type of website, not a complete website or protocol.

REFERENCES

[1] "telecomhall." [Online]. Available: http://www.telecomhall.com/
[2] "Tm forum inform." [Online]. Available: https://inform.tmforum.org/
[3] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, p. 175, 1992.
[4] B. Anderson and D. Mcgrew, "Machine learning for encrypted malware traffic classification," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 17*, 2017.
[5] S. Ansari, S. G. Rajeev, and H. S. Chandrashekar, "Packet sniffing: a brief introduction," *IEEE Potentials*, vol. 21, no. 5, pp. 17–19, Dec 2002.
[6] R. Bendrath and M. Mueller, "The end of the net as we know it? deep packet inspection and internet governance," *New Media & Society, volume = 13, number = 7, pages = 1142-116*, 2011.
[7] N. Boudriga, *Security of mobile communications*. CRC Press/Taylor Francis, 2010.
[8] A. Bremler-Barr, S. T. David, Y. Harchol, and D. Hay, "Leveraging traffic repetitions for high-speed deep packet inspection," *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015.
[9] R. T. El-Maghraby, N. M. A. Elazim, and A. M. Bahaa-Eldin, "A survey on deep packet inspection," in *2017 12th International Conference on Computer Engineering and Systems (ICCES)*, Dec 2017, pp. 188–197.
[10] A. Ghaderi and V. Athitsos, "Selective unsupervised feature learning with convolutional neural network (s-cnn)," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 2486–2490.
[11] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, July 1998.
[12] K. Ihalagedara, R. Kithuldeniya, S. Weerasekara, and S. Deegalla, "Feasibility of using machine learning to access control in squid proxy server," *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*, 2015.
[13] J. Jiang, J. Jiang, B. Cui, and C. Zhang, "Tencentboost: A gradient boosting tree system with parameter server," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, April 2017, pp. 281–284.
[14] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane, "Automated design of both the topology and sizing of analog electrical circuits using genetic programming," *Artificial Intelligence in Design 96*, p. 151170, 1996.
[15] B. Li, E. Zhou, B. Huang, J. Duan, Y. Wang, N. Xu, J. Zhang, and H. Yang, "Large scale recurrent neural network on gpu," in *2014 International Joint Conference on Neural Networks (IJCNN)*, July 2014, pp. 4062–4069.
[16] M. Lotfollahi, R. S. H. Zade, M. J. Siavoshani, and M. Saberian, "Deep packet: A novel approach for encrypted traffic classification using deep learning," *CoRR*, vol. 1709.02656, 2017.
[17] M. Mishra and M. Srivastava, "A view of artificial neural network," in *2014 International Conference on Advances in Engineering Technology Research (ICAETR - 2014)*, Aug 2014, pp. 1–3.
[18] R. Oppliger, "Internet security: firewalls and beyond," *Communications of the ACM*, vol. 40, no. 5, p. 94, Jan 1997.
[19] T. J. Parvat and P. Chandra, "Performance improvement of deep packet inspection for intrusion detection," in *2014 IEEE Global Conference on Wireless Computing Networking (GCWCN)*, Dec 2014, pp. 224–228.
[20] T. R. Peltier and J. Peltier, *Complete guide to CISM certification*. Auerbach, 2007.
[21] J. Quinlan, "Simplifying decision trees," *International Journal of Human-Computer Studies*, vol. 51, no. 2, p. 497510, 1999.
[22] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, p. 147, Jan 2002.
[23] C. L. Van Jacobson and U. o. C. B. C. Steven McCanne, all of the Lawrence Berkeley National Laboratory. Tcpdump user commands. [Online]. Available: http://www.tcpdump.org/manpages/tcpdump.1.html

[24] C. Yan, "Innovative applications of artificial intelligence," *Engineering Applications of Artificial Intelligence*, vol. 3, no. 2, p. 166167, 1990.
[25] H. Yu, Y. Zhao, G. Xiong, L. Guo, Z. Li, and Y. Wang, "Poster: Mining elephant applications in unknown traffic by service clustering," *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS 14*, 2014.
[26] H. Zimmermann, "Osi reference model - the iso model of architecture for open systems interconnection," *IEEE Transactions on Communications*, vol. 28, no. 4, pp. 425–432, April 1980.