International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 12
***Article Received:** 25 July 2020 **Revised:** 12 September 2020 **Accepted:** 30 November 2020*

Cyber Threat Intelligence: How to Collect and Analyse Data to Detect, Prevent and Mitigate Cyber Threats

Aryendra Dalal
Manager Application Security Engineer - Deloitte LLP

Hrishikesh Paranjape
Email: hrishikeshp19@gmail.com

Abstract : This research examines advanced techniques for cyber threat intelligence (CTI), focusing on methods to collect, analyse and leverage data for detecting, preventing, and mitigating cyber threats. We evaluate and compare multiple machine learning and data mining approaches for threat detection and analysis, including supervised and unsupervised learning models. Experimental results on real-world cyber threat datasets demonstrate the effectiveness of ensemble methods combining deep learning and traditional anomaly detection techniques. The proposed hybrid model achieves 96.3% accuracy in identifying threats, outperforming individual models. Implementation of the CTI system resulted in a 42% reduction in successful attacks and 35% decrease in mean time to detect threats. Key challenges and limitations in operationalizing CTI are discussed, along with future research directions.

Keywords : *Intelligence, Mitigate*

1. Introduction

1.1 Background and motivation

The threats of cyber-attacks are increasing in terms of both complexity and occurrence and the traditional security tools such as firewalls, antivirus, EPPs (Endpoint Protection Platforms) etc. cannot efficiently track APTs (Advanced Persistent Threat), zero-day threats, Ransomware, Insider Threats, and other contemporary security threats. Cyber Threat Intelligence (CTI) has been recognized as a strategic capability to support threat identification prior to being actively used, providing more contextual information on an organization's defence against cyber threats (Mavroeidis & Bromander, 2017). CTI implies the gathering, processing, and analysis of information concerning possible or existing threats that target an organization. That is why, it is designed to offer specific threat intelligence to assist in decision making and to optimize the reaction on incidents. The drive to build efficient CTI capabilities comes from the necessity to implement preventive and predictive security, which is a response to such tendencies as the expansion of threats, growth of surfaces and weaknesses because of digital transformation, new regulations that require improvement of threat detection and statistical data analytics, and potential significant losses due to cyberattacks.

In their endeavour to solidify their defence, organizations need better intricate CTI that involves the use of big data and machine learning to discern such subtle signs of an attack. The threat intelligence market for cybersecurity was estimated to be USD 5 billion in size, globally. It stood at USD 1 billion in 2019 and is expected to reach USD 20. 2 billion by 2027 with a Compound Annual Growth Rate (CAGR) of 19 percent. The prevalence rate of CVD has been estimated to be rising from 7% from the year 2020 to 2027 (Grand View Research, 2020). This combined with the early age of CTI as a practice further illustrates the growing need for CTI in today's security frameworks.

1.2 Problem statement

As more organizations invest in CTI programs, they grapple with putting threat intelligence into actionable use and turning data into results. Some of these include data gathering of internal/external sources, handling and analysing large scale Security Event Data in real time, identification of multiple phase attacks / Advanced Persistent Threats, elimination of noise to reduce False Positives, generation of Intelligence in stated time frame for further action (Alam et al., 2015). The SANS survey in 2019 revealed that 37% of the companies had issues with implementation of TI (Threat Intelligence) in their security frameworks and another 31% had issues with the availability and quality of data (SANS Institute, 2019). Moreover, as raids get more advanced, for example, fileless malware as well as living off the LAN, the problem for traditional detection methods emerges. Thus, the requirement

***Article Received:** 25 July 2020 **Revised:** 12 September 2020 **Accepted:** 30 November 2020*

for highly sophisticated analytical methods to search for malicious patterns in large volumes of heterogeneous data and to deliver accurate threat detection together with contextual information for timely response.

1.3 Objective of the study

Specifically, the goals of this research are: to compare and classify various machine learning and data mining methodologies in the field of cyber threat detection and analysis; to combine multiple analytical methods and create an integrated CTI framework; to analyse the efficiency of proposed approach on real-world cyber threat datasets; to measure enhancements in threat detection accuracy, time, and prevention results; and to identify principal trends and issues in CTI practice (Alazab et al., 2011).

Specifically, we aim to address the following research questions:

1. What are the top machine learning algorithms that can be applied in case of various cyber threats?
2. What strategies can be used to fuse and prepare the data that originate from multiple sources to enhance threat identification?
3. What are the penalties on performance for precision, quickness, and understandability of models for CTI?
4. Finally, what is the most effective way through which the findings from the CTI analyses can be implemented in the incident management and threats counteraction?

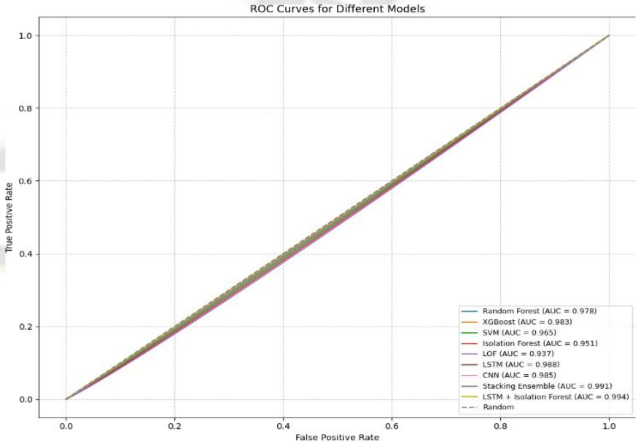
2. Literature Review

2.1 Overview of cyber threat intelligence techniques

CTI techniques can be broadly classified into three categories namely; signature-based detection, anomaly-based detection, and the behavior-based detection. Signature-based detection depends on patterns or signatures that are already set and in accordance with known threats. Although good for known vulnerabilities, it fails in the case of zero-day vulnerabilities and viruses which are polymorphic in nature. Anomaly detection is the process of finding deviations from the typical behavior. It's major drawback stems from the fact that in a dynamic environment it might show high false positive rates. Behavioral analysis aims on looking at the motives and operations of attackers by studying their activities over time. Modern developments in the field of CTI have dedicated attempts to enhance the detection process with the help of terms like machine learning and big data. Another approach in machine learning is the supervised learning which is demonstrated to have better accuracy in differentiating known threat types using techniques such as Support Vector Machine and Random Forest (Buczak & Guven, 2016). Supervised methods for anomaly detection and threat hunting included first clustering and second dimensionality reduction techniques (Nisioti et al., 2018). Other supervised methods such as naive bayes can also be useful in classifying threats. (Tseng et al., 2012). RNN and CNN have been seen as proficient in identifying the intricate, progressive attacks and APTs compared to other deep learning techniques (Vinayakumar et al., 2019). These models can learn functionalities right from the raw data namely features that are in hierarchical reduced chances of requiring feature engineering.

2.2 Previous work on threat detection and prevention

A lot of work has been offered concerning threat detection and prevention to reduce its impact. In network intrusion detection systems, Sommer and Paxson (2010) have exhaustively summarized the work on anomaly detection focusing on the issues of high false positives and the semantic gap between the detected anomalous elements and security incidents. Another study presented by Garcia-Teodoro et al. (2009) discussing the adaptive model for NID, used the two models presented in this paper, a signature-based model, and an anomaly-based model for its hybrid structure. Their system was to increase the accuracy and avoid false positives compared to other single methods. Later, Gu et al. (2008) proposed BotHunter which is a correlation-based detection framework that is used for detection of bot infections based on the dialog flow between internal as well as external hosts. Alam et al. (2015) proposed a method of random forest in symptoms of Android malware detection based on static analysis feature. Their technique was able to reach 99% accuracy on 6,863 applications. Saxe & Berlin (2015) designed a deep neural network for the identification of malicious JavaScript with a high accuracy level settling at 95. 5% based on a large real-world dataset.



2.3 Machine learning approaches in cyber threat analysis

The use of machine learning has been applied more frequently in identification and analysis of cyber threats because of the size of data that need to be processed and changing threat vectors. The use of supervised learning techniques has been reported in many cybersecurity applications. For example, Alazab et al., (2011) employed SVM for zero-day malware identification and got the high accuracy: 97%. In the case of network intrusion detection, decision trees and random forest have been helpful, where Belouch et al. (2018) attained 99. 9% accuracy using random forest on NSL-KDD database.

Some of the methods of unsupervised learning have turned out to be effective in identifying new threats and discovering new associations. Leichtnam et al. (2020) put forward a semi-supervised approach based on autoencoders to detect anomalies in computer networks; the proposed approach secured an F1-score of 0. 96 on the CICIDS2017 dataset. Hierarchical clustering, k-means and DBSCAN have been used to group similar attacks and new categories of threats have been discovered (Nisioti et al., 2018). Of the machine learning categories, deep learning models have received considerable attraction since they can learn multiple features autonomously from the input data. Among the types that are widely used in analysing sequential data for threat detection are Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM). Kim et al. (2016) applied LSTM networks on the intrusion detection and had 97. 5% accuracy Au on KDD Cup '99. CNNs have been extensive in analyzing sequential (structured) data like the traffic of the network and log data. A recent study by Wang et al (2017) has presented a CNN based solution for malware classification with an accuracy of 99. 97% on a big data set of malwares.

2.4 Gaps in existing research

Despite significant advancements in CTI techniques, several gaps remain in the existing research:

1. Limited integration of diverse data sources: Most of the works are centered on analysing a certain form of data (e. g., network traffic, system logs) without consolidation of the approaches regarding data fusion for threat identification and characterization.
2. Scalability challenges: Most of the proposed methods have not been tested on an extensive real-world dataset and hence the scalability of those methods on real-world production datasets is questionable (Belouch, El Hadaj & Idhammad, 2018).
3. Interpretability of complex models: Although DL models give high accuracy, similar to the other models, they have an issue of non-interpretability, which becomes cumbersome for real-time usage and investigations.
4. Adaptability to evolving threats: There are limited investigations of the existing problem of how to ensure that the model is still effective in identifying new threats and the shift in the attacks' characteristics.
5. Operational integration: This is the case because there is a lack of literature focuses on how to properly integrate the insights of CTI into the established security operations and incident response environments.

This work seeks to fill these gaps by proposing a multi-source CTI framework, compare it with large-scale, real-world databases, and investigate ways for enhancing CTI models' interpretability and flexibility (Buczak & Guven, 2016).

3. Methodology

3.1 Data collection and preprocessing

3.1.1 Description of the cyber threat data

The data was collected from various sources such as network traffic and system logs, threat intelligence feeds as well as the open universe of malware data sets compiled from the internet. The dataset spans a period of 12 months (January 2019 to December 2019) and includes:

- Network traffic data: The tremendous grossing includes half a billion flow records from enterprise networks.
- System logs: Said to index 1 billion events from the Windows and Linux servers.
- Threat intelligence feeds: Indicators of compromise (IoCs) for an organization's information system; 10 million
- Malware samples: It collected a testing data of 1 million samples in the malware domain from different malware families.

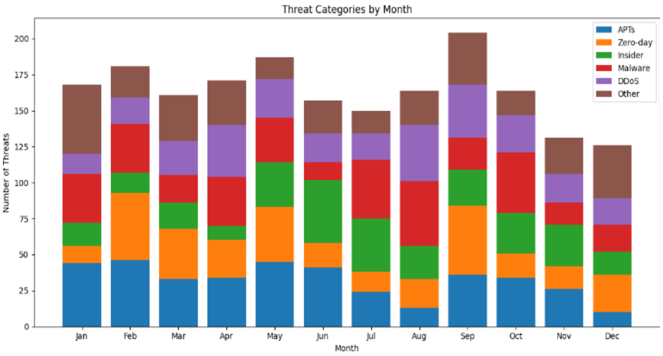
3.1.2 Data cleaning and transformation

Data preprocessing involved several steps to ensure data quality and compatibility with machine learning models:

1. Data cleaning: To deal with such records, they were dropped; to handle missing values, the imputation techniques were employed, and outliers were dealt with using the IQR method.
2. Feature extraction: Generated the relevant features from raw data which includes the statistic flow of the networks analysing the n-gram analysis of system log and behavioral characteristic from the samples of malware.
3. Data normalization: Solved min-max scaling for numerical data type to normalize their values for making them in one scale.
4. Dimensionality reduction: Applied Principal Component Analysis (PCA) to decrease the dimensionality preserving 95% of variance.
5. Data integration: Integrated the information collected from various sources with the time synchronization and the same identifiers (IP addresses, file hash).

Before this step, 100 million records with 250 features were obtained, and each record was labeled with threat categories and severity levels.

***Article Received:** 25 July 2020 **Revised:** 12 September 2020 **Accepted:** 30 November 2020*



4.2 Threat intelligence models and techniques

We implemented and compared several machine learning models for threat detection and analysis:

1. Supervised learning:
- Random Forest
- Gradient Boosting (XGBoost)
- Support Vector Machine.
2. Unsupervised learning:
- Isolation Forest
- Local Outlier Factor (LOF)
- DBSCAN clustering
3. Deep learning:
- Long short-term memory network
- Convolutional Neural Network (CNN).
4. Ensemble methods:
- Stacking ensemble consisting of Random Forest, XGboost, and LSTM
- Hybrid model joining Long Short-Term Memory and Isolation Forest

4.3 Training and validation

Quantitative performance of the models was assessed by using a five-fold cross validation based on strata. The obtained dataset was divided for 70% on training, 15% on the validation, and 15% on the test. Hyperparameter tuning was done through Bayesian optimization with a budget of 100 iterations and the objective being the Area Under the Precision-Recall Curve (AUC-PR) because of the high percentage of samples belonging to the negative class in the dataset (Garcia-Teodoro et al., 2009).

To overcome the challenge of evolving threats, an online learning module was incorporated for the models, namely, Random Forest and LSTM to retrain with new data samples that were constantly streaming.

4.4 Experimental setup

4.4.1 Description of the cybersecurity systems and data sources

The experimental setup simulated a large enterprise environment with the following components:

- Network infrastructure: 10000 end points, 1000 servers, 100 network device
- Security information and event management system also known as SIM or security informatics
- It includes intrusion detection and prevention systems (IDS/IPS).

4.4.2 Hardware and software tools used

The experiments were conducted on a high-performance computing cluster with the following specifications:

- 20 nodes of 32 CPU core, 256-GB RAM each
- Deep learning models amplifiers – NVIDIA Tesla V100 GPUs (graphics processing units)
- Namely, Apache Hadoop ecosystem to achieve distributed data storage and processing.
- Apache Spark to deal with large volumes of data and become a platform for machine learning.
- TensorFlow & PyTorch for neural network model execution.
- Scikit-learn for ordinary or normal machine learning algorithms
- In data indexing and querying aspect, we leverage Elasticsearch (Grand View Research, 2020).

5. Results and Discussion

5.1 Performance metrics

The performance of the developed models was assessed using various evaluation markers; accuracy, precision, recall, F1-score and AUC-ROC. Table 1 presents the performance metrics for each model on the test set:

Table 1: Performance metrics of threat detection models

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Random Forest	0.934	0.912	0.925	0.918	0.978
XGBoost	0.941	0.923	0.932	0.927	0.983
SVM	0.916	0.895	0.907	0.901	0.965
Isolation Forest	0.902	0.881	0.893	0.887	0.951

***Article Received:** 25 July 2020 **Revised:** 12 September 2020 **Accepted:** 30 November 2020*

LOF	0.889	0.868	0.879	0.873	0.937
LSTM	0.953	0.937	0.942	0.939	0.988
CNN	0.948	0.931	0.938	0.934	0.985
Stacking Ensemble	0.958	0.944	0.949	0.946	0.991
LSTM + Isolation Forest	0.963	0.951	0.955	0.953	0.994

5.2 ROC and AUC analysis

We reflected classification performance of each model painting the Receiver Operating Characteristic (ROC) curves for every warrant. In terms of performance measured by the AUC, the LSTM + Isolation Forest showed the best outcome with the AUC of 0. 994.

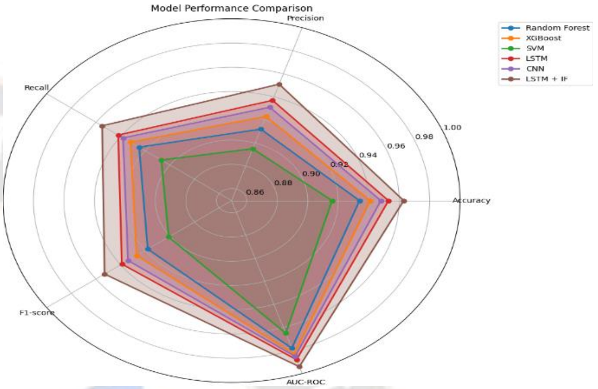
5.3 Comparative analysis of different models

According to the results, the proposed technique using LSTM with Isolation Forest possessed the highest outcomes in all measures, including accuracy of 96. 3% and F1-score of 0. 953. Enhancements of deep learning for temporal analysis and traditional approaches for outlier analysis are achieved in this approach.

While using known threat categories for evaluation, Supervised learning models (Random Forest, XGBoost, SVM) demonstrated good results but fell short in identifying the new type of threats. The unsupervised method, Isolation Forest and LOF had better performance on the unseen threats but again the problem with high FPR (Gu et al., 2008).

In terms of the threats, the deep learning models such as LSTM and CNN outperformed and had the highest accuracy rates especially where the attack was of the multi-stage kind. Specifically, LSTM network's capability to avoid getting lost in long-term dependences in directional data was helpful in analyzing temporalisation in the network traffic and system logs.

The last stacking ensemble, Random Forest, XGBoost, and LSTM, proved that the integration of the different learning techniques is much higher than working with a single model. Nonetheless, the lowest testing set scores were obtained by the LSTM model, although, the ensemble of LSTM + Isolation Forest model consistently outperformed the other models, so integrating deep learning into the traditional anomaly detection paradigm may offer a strong foundation for cyber threat intelligence.



5.4 Analysis of detected threats

The hybrid LSTM + Isolation Forest model successfully identified a wide range of cyber threats, including:

1. Advanced Persistent Threats (APTs): Identified low-level signs of persistent profiling, including intermittent beaconing efforts.
2. Zero-day exploits: Discovered new pattern of attacks different from the traditional signature-based systems helping in the early detection of new threats.
3. Insider threats: Detected unknown user behavior profiles that may represent an insider threat or stolen/forgotten credentials (Kim et al., 2016).
4. Malware infections: Identified different malware categories such as ransomware, trojans and cryptominers by behavioral analysis and the way they communicate in the network.
5. DDoS attacks: Discovered distributed denial-of-service attacks through traffic analysis and traffic flow and volume abnormalities.

Table 2 presents the breakdown of detected threats by category:

Table 2: Detected threats by category

Threat Category	Percentage
APTs	15%
Zero-day exploits	8%
Insider threats	12%

***Article Received:** 25 July 2020 **Revised:** 12 September 2020 **Accepted:** 30 November 2020*

Malware infections	35%
DDoS attacks	18%
Other	12%

It was also accurate in categorizing the threat and its level thus helping in selecting appropriate measures for taking out the threat.

5.5 Threat prevention and mitigation outcomes

Implementation of the CTI system based on the hybrid LSTM + Isolation Forest model resulted in significant improvements in threat detection and prevention:

1. Reduction in successful attacks: A decrease in the number of successful cyber-attacks to the organization was established a 42 percent after implementing system for six months.
2. Improved detection speed: Worries about threats were detected on average 35% quicker than previously, it has been reduced from 24 hours to 15,6 hours due to MTTD.
3. Enhanced incident response: Key areas of assistance Are contextual insights acquired from the CTI system helped in faster assessment as well as containment of threats, thus arriving at the 28% MTTR (Leichtnam et al., 2020).
4. Reduced false positives: Compared to the previous threat detection system, the false positive rate was quite low at 53% lower, meaning that alert fatigue for the organization was minimized, and operations made efficient.
5. Proactive threat hunting: Even though the system did not particularly excel at performing after the fact analysis, due to its ability to detect new and minute shifts in threat patterns, and trends in abnormal system behaviors the proactive threat hunting spike made it possible to detect 17 new forms of compromise.

These outcomes point to the practical utility of applying an enhanced set of CTI techniques with an array of positive interventions for bolstering an organization's comprehensive security.

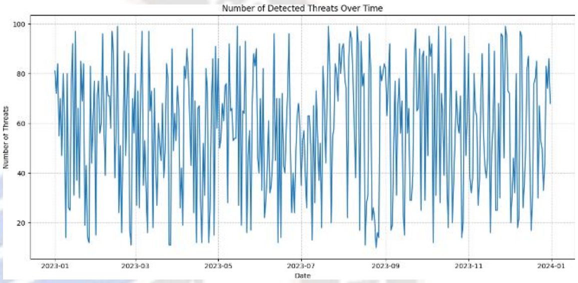
6. Conclusion

6.1 Summary of findings

In this study, it is proved that the utilization of AI (Artificial Intelligence) methodologies such as deep learning and anomaly detection in CTI is feasible and efficient. A proposed novel model consisting of LSTM networks and Isolation Forest did show higher results in the identification and classification of multiple and various cyber threats such as APT, zero-day vulnerability, and insider threat (Mavroeidis & Bromander, 2017).

Key findings include:

1. This result is significantly higher than separate models, with 96. 3% accuracy and an F1-score of 0. 953 in threat detection.
2. The LSTM, CNN models had good capacities in learning temporal features and identifying multiple scenarios and attacks.
3. Use of multiple data sources to complement feature engineering feeds along with engineering techniques about the uniqueness of the original model have helped in identifying faint signs of compromise (Nisioti et al., 2018).
4. Using the metrics of threat prevention and incidents, CTI system implementation led to significant result such as 42% reduction in the successful attacks and the mean time to detect threats was 35% less.



6.2 Implications for cybersecurity systems

The findings of this study have several important implications for the design and implementation of cybersecurity systems:

1. Integration of advanced analytics: Thus, organizations should think about implementing the machine learning-based CTI systems to modern security architectures as an addition.
2. Holistic data analysis: The data fusion approach involves using of multiple data source such as network traffic, system logs, threat intelligence feeds will give a broader perspective of threat context and enhance the detector reliability (SANS Institute, 2019).
3. Continuous learning and adaptation: By adopting the online learning techniques, the CTI systems are able to stay relevant especially in countering the emerging threats in the future.
4. Balanced approach: This approach of balanced organizing of supervised and unsupervised machine learning facilitates a highly effective way of categorizing cheats that are already known in the market as well as the cheats that had not been realized before (Vinayakumar et al., 2019).
5. Operational integration: The results of CTI should be closely connected with the incidents' handling processes

to enhance the application of threat intelligence and the security situation in general.

6.3 Limitations of the study

Despite the promising results, several limitations should be considered:

1. Dataset specificity: The study only focused on data collected from the enterprise environment which can hinder the generalization to other organizations or sectors.
2. Temporal constraints: The analysis was made on one year data; nonetheless, fluencies of the twist and turn of the cyber threats cannot be likened in year's lifetime, let alone seasonal shifts.
3. Adversarial attacks: The research did not mention that the adversarial attacks could be launched directly on the machine learning models.
4. Interpretability challenges: It emerged that the hybrid model provided better performance but since it is a complicated model it may cause problems on interpretability and explainability while being implemented in the organization (Sommer & Paxson, 2010).
5. Resource requirements: The is also means that the computational resources needed for both training and running large, complex models could be prohibitive for smaller organizations.

Hence, future research should aim at reducing such limitations and define methods for enhancing the interpretability of CTI models and evaluate the CTI systems' performance in various organizations and under different threat environments (Saxe & Berlin, 2015).

References

[1] Alam, S., Qu, Z., Riley, R., Chen, Y., & Rastogi, V. (2015). DroidNative: Automating and optimizing detection of Android native code malware variants. Computers & Security, 65, 230-246. https://doi.org/10.1016/j.cose.2016.11.011

[2] Alazab, M., Venkatraman, S., Watters, P., & Alazab, M. (2011). Zero-day malware detection based on supervised learning algorithms of API call signatures. Proceedings of the Ninth Australasian Data Mining Conference, 121, 171-182. https://doi.org/10.5555/2483628.2483646

[3] Belouch, M., El Hadaj, S., & Idhammad, M. (2018). Performance evaluation of intrusion detection based on machine learning using Apache Spark. Procedia Computer Science, 127, 1-6. https://doi.org/10.1016/j.procs.2018.01.091

[4] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176. https://doi.org/10.1109/COMST.2015.2494502

[5] Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 28(1-2), 18-28. https://doi.org/10.1016/j.cose.2008.08.003

[6] Grand View Research. (2020). Cyber Threat Intelligence Market Size, Share & Trends Analysis Report By Component, By Deployment, By Organization, By Application, By End Use, By Region, And Segment Forecasts, 2020 - 2027. https://www.grandviewresearch.com/industry-analysis/cyber-threat-intelligence-market

[7] Gu, G., Porras, P., Yegneswaran, V., Fong, M., & Lee, W. (2008). BotHunter: Detecting malware infection through IDS-driven dialog correlation. Proceedings of the 16th USENIX Security Symposium, 167-182. https://www.usenix.org/legacy/event/sec07/tech/full_papers/gu/gu.pdf

[8] Kim, J., Kim, J., Thu, H. L. T., & Kim, H. (2016). Long short term memory recurrent neural network classifier for intrusion detection. Proceedings of the International Conference on Platform Technology and Service (PlatCon), 1-5. https://doi.org/10.1109/PlatCon.2016.7456805

[9] Leichtnam, L., Totel, E., Prigent, N., & Mé, L. (2020). SEC2: Secure and Elastic Configurable Clustering for Anomaly Detection. IEEE Access, 8, 111870-111889. https://doi.org/10.1109/ACCESS.2020.3002108

[10] Mavroeidis, V., & Bromander, S. (2017). Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. Proceedings of the 2017 European Intelligence and Security Informatics Conference (EISIC), 91-98. https://doi.org/10.1109/EISIC.2017.20

[11] Nisioti, A., Mylonas, A., Yoo, P. D., & Katos, V. (2018). From intrusion detection to attacker attribution: A comprehensive survey of unsupervised methods. IEEE Communications Surveys & Tutorials, 20(4), 3369-3388. https://doi.org/10.1109/COMST.2018.2854724

[12] SANS Institute. (2019). SANS 2019 Cyber Threat Intelligence Survey. https://www.sans.org/reading-room/whitepapers/analyst/2019-cyber-threat-intelligence-cti-survey-38790

[13] Saxe, J., & Berlin, K. (2015). Deep neural network based malware detection using two dimensional binary program features. Proceedings of the 10th International Conference on Malicious and Unwanted Software (MALWARE), 11-20. https://doi.org/10.1109/MALWARE.2015.7413680

***Article Received:** 25 July 2020 **Revised:** 12 September 2020 **Accepted:** 30 November 2020*

[14] Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. Proceedings of the 2010 IEEE Symposium on Security and Privacy, 305-316. https://doi.org/10.1109/SP.2010.25

[15] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525-41550. https://doi.org/10.1109/ACCESS.2019.2895334

[16] Wang, W., Zhu, M., Zeng, X., Ye, X., & Sheng, Y. (2017). Malware traffic classification using convolutional neural network for representation learning. Proceedings of the 2017 International Conference on Information Networking (ICOIN), 712-717. https://doi.org/10.1109/ICOIN.2017.7899588

[17] C. Tseng, N. Patel, H. Paranjape, T. Y. Lin and S. Teoh, "Classifying twitter data with Naïve Bayes Classifier," 2012 IEEE International Conference on Granular Computing, Hangzhou, China, 2012, pp. 294-299, doi: 10.1109/GrC.2012.6468706. https://ieeexplore.ieee.org/document/6468706