KIPS Trans. Comp. and Comm. Sys.
Vol.10, No.2 pp.39~46
ISSN: 2287-5891 (Print), ISSN 2734-049X (Online)

https://doi.org/10.3745/KTCCS.2021.10.2.39

A Study on Malicious Code Detection Using Blockchain and Deep Learning

Deok Gyu Lee†

ABSTRACT

Damages by malware have recently been increasing. Conventional signature-based antivirus solutions are helplessly vulnerable to unprecedented new threats such as Zero-day attack and ransomware. Despite that, many enterprises have retained signature-based antivirus solutions as part of the multiple endpoints security strategy. They do recognize the problem. This paper proposes a solution using the blockchain and deep learning technologies as the next-generation antivirus solution. It uses the antivirus software that updates through an existing DB server to supplement the detection unit and organizes the blockchain instead of the DB for deep learning using various samples and forms to increase the detection rate of new malware and falsified malware.

Keywords : Malicious Code, Code Detection, Blockcahin, Deep Learning

블록체인과 딥러닝을 이용한 악성코드 탐지에 관한 연구

이 덕 규†

요 약

최근 맬웨어에 의한 피해가 증가하고 있다. 기존의 시그니처 기반 안티 바이러스 솔루션은 제로 데이 공격 및 랜섬웨어와 같은 새로운 위협에 취약하다. 그럼에도 많은 기업은 문제점을 인식하고, 다중 엔드 포인트 보안 전략의 일부로 서명 기반 안티 바이러스 솔루션을 유지하고 있다. 본 논문에서는 차세대 안티 바이러스 솔루션으로 블록 체인과 딥 러닝 기술을 이용한 솔루션을 제안한다. 기존 DB 서버를 통해 업데이트되는 바이러스 백신 소프트웨어를 사용하여 탐지 유닛을 보완하고, 다양한 샘플과 형태를 사용하여 딥 러닝 용 DB 대신 블록 체인을 구성하여 신규 악성 코드 및 위조 악성 코드 탐지율을 높이는 방법을 제안한다.

키워드 : 악성코드, 코드 탐지, 블록체인, 딥러닝

1. Introduction

KEnterprise users use notebook PCs, tablets, smartphones, and many storage units in addition to desktop PCs. The advanced malware takes advantage of vulnerabilities of such endpoint devices to threaten enterprise security. According to statistics by Kaspersky, the number of detected malicious files known as a backdoor in 2018 increased by 44%, and the size of ransomware also increased by 43%. In other words, about 1/3 (30.01%) of all computers experienced one or more online malware attacks in 2018.[1] Therefore, in this paper, the shortcomings of the existing malicious code countermeasure solutions are supplemented using deep learning and blockchain to solve. With regard to data collection, a problem with deep running systems, we propose a system that can continuously collect and utilize data using blockchain. Chapter 2 reviews previous studies, Chapter 3 describes the proposed method, and Chapter 4 presents the conclusion.

2. Related Works

A state-of-the-art survey of malware detection approaches using data mining techniques

※ This work was partly supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea Government(MSIT) (No. 2020-0-00326, Development of smart port application platform by real-time tracking of logistics information based on blockchain)
† 종신회원 : 서원대학교 정보보안학과 조교수
Manuscript Received : December 2, 2020
Accepted : December 8, 2020
* Corresponding Author : Deok Gyu Lee(deokgyulee@gmail.com)

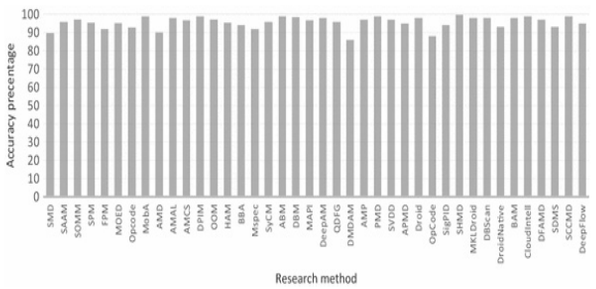

Fig. 1. Accuracy Factor for Selected Approaches in Malware Detection[5]

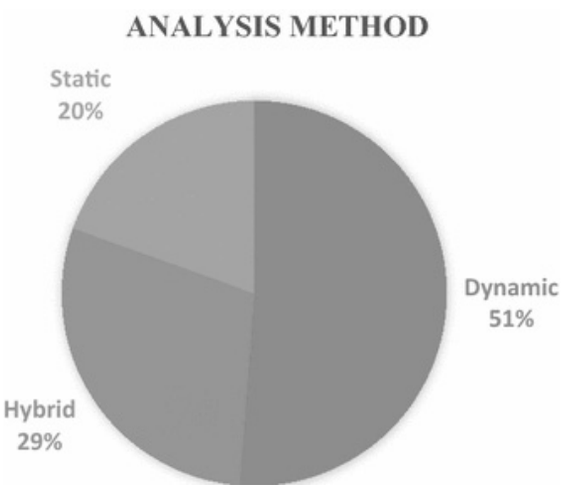

Fig. 2. The Data Analysis Methods in the Selected Articles[5]

Fig. 1 shows the main case study diagram of each research in malware detection. As shown, the recent researches have considered android smartphones to analyze malware detection approaches with 40%. The symbolic code aggregation case studies in windows-based platform has 23%, the pattern mining has 11%, the system calls has 8% usage in malware detection.[5]

Also, Fig. 2 shows the data analysis methods percentage in terms of static, dynamic and hybrid analysis in selected research. The most data analysis methods have used dynamic analysis with 51%, the hybrid analysis has 29% and the static analysis has 20% usage. The 30% of the signature-based approaches have used the dynamic data analysis. The 65% of the behavior-based malware detection approaches have used the dynamic data analysis method.[5]

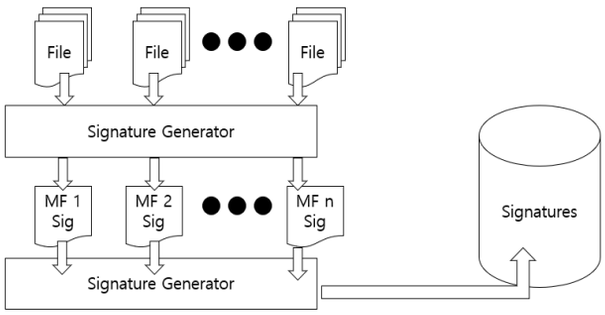

Fig. 3. Signature Generation Process[7]

2.1 Study of Real-Time Malware Detection in Intrusion Detection System

The intrusion detection system matches the transferred network packet with the defined detection signature, as shown in Fig. 3 and performs the task specified by the detected signature depending on the result. The detection generation algorithm defines the detection signature in advance, and the match algorithm signature defines the detection signature match method for monitoring of the packets. There are already cases of including the signature, which monitors infection attempts by the malware, in the intrusion detection system. However, the existing methods only monitor the packet behaviors and thus cannot block the transfer of malicious programs. The detection signature generation algorithm generates the detection signature to determine if a payload is part of a malicious program. It extracts the signature for each known malware family and converts them into a character string to generate the detection signature.[7]

2.2 Signature-Based Detection Methodology

The signature-based detection method is called Misuse Detection, which is the methodology that uses the signature or pattern to detect all known attacks that match. Although this method has the advantage of accurately detecting all known attacks defined with patterns, it generally has the difficult of detecting cyber threats with almost unknown attacks against control systems. There is a more serious problem of not being able to detect new types of attacks such as Zero-day attack until the detection patterns are continuously developed through the detailed analysis and applied to the system.[8]

Attack detection in water distribution systems using machine learning

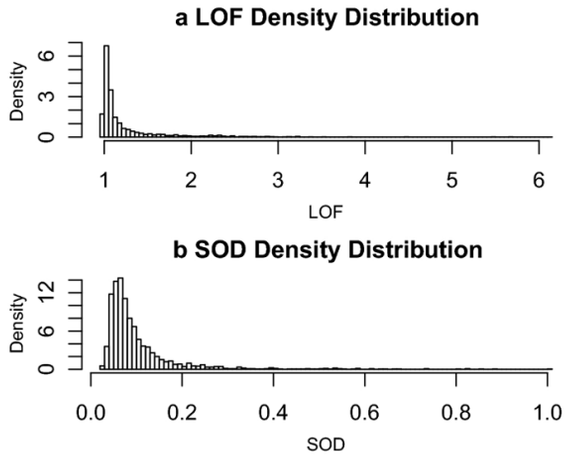

Fig. 4. Density Distribution of LOF and SOD[13]

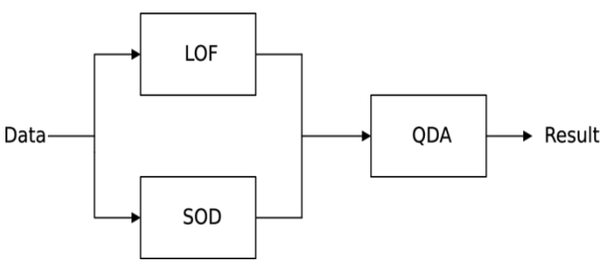

Fig. 5. Proposed Ensemble Technique[13]

The proposed ensemble technique combines both SOD and LOF using QDA as shown in Fig. 5. As can be seen from the figure a datapoint is first run through both algorithms in parallel and the degree to which the value is an outlier is calculated using both (2) and (4). This density based phase of the proposed ensemble technique outputs two values, the LOF and SOD respectively. As discussed previously, an LOF and SOD value of much larger than one and zero respectively means that the datapoint is an anomaly. This is also evident when looking at the density distribution shown in (3). LOF will outperform SOD when run on low dimensional data and the opposite is true when considering high dimensional data. The proposed technique incorporates both of these values in order to create a more robust algorithm that leverages the advantages of the baseline algorithms. To find a more complex decision boundary, QDA is introduced to find a model that uses both values to classify the data. This was chosen over popular methods such as bagging and boosting normally used in ensemble classifiers in order to

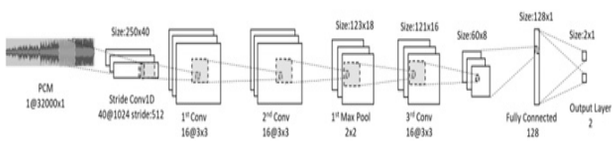

Fig. 6. CNN Algorithm Structure[14]

leverage the classification power of QDA. These traditional methods normally rely on the classification outputs of each of the algorithms which are then combined to produce more accurate results. The advantage of QDA is that it uses the outlier values produced by each algorithms to draw its own independent decision boundary. The density distribution of the data means that QDA will be able to be able to produce very accurate results.[13]

2.3 Utilization of Deep Learning

Facebook announced a facial recognition service in June 2015. It scans photos and classifies them based on specific people in the photos. When sharing a group photo, it finds the similarities between images using the convolution neural network (CNN) and the sharing photos between friends without email and snapshots and assigns the weight factor to the similarities. It then learns by changing the weight factor of similarities to improve the accuracy of the matching of two images by using the CNN algorithm. Google's Alpha Go is another example of using DNN. It uses the neural network of 12 layers and more than 1 million neurons to learn the moves of go games. It does not just mimic the movement patterns and learns and coordinates the data in the neural network to generate strategies autonomously.[15]

2.4 CNN(Convolutional Neural Network)

CNN is a imitation of a human optic nerve. Although previous methods were learned by extracting knowledge from data, CNN's structure is to identify patterns of these features by extracting data into feature. This CNN algorithm is conducted through the Convolution process and the Pooling process. The Convolution Layer and the Pooling Layer are compounded to create algorithm. The use of this CNN is usually used for Information Extraction, Sentence Classification and Face Recognition.

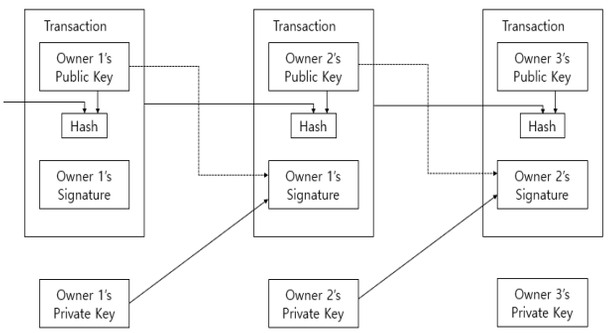

Fig. 7. Blockchain Principle

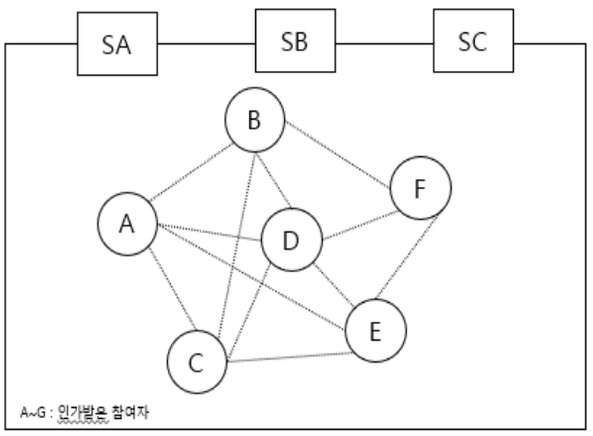

Fig. 8. Consortium Blockchain Principle

2.5 Overview of Blockchain

The blockchain was first introduced with the appearance of Bitcoin developed by someone under the pseudonym Satoshi Nakamoto in 2008. The technology requires all participants to verify the feasibility of the transaction when a new transaction occurs. The approved transaction is accepted as the newly generated block and chained to existing blocks.

A transaction is completed when the participants store the copy of the updated blockchain in distribution. The blockchain is a type of distributed ledger with which all participants of the network share the transaction records.[9]

2.6 Consortium BlockChain

A consortium blockchain is a quasi-centralized blockchain formed by multiple organizations as the co-subjects. Like private blockchains, only the authorized targets, each with different permissions, can participate in it, and the targets to disclose the

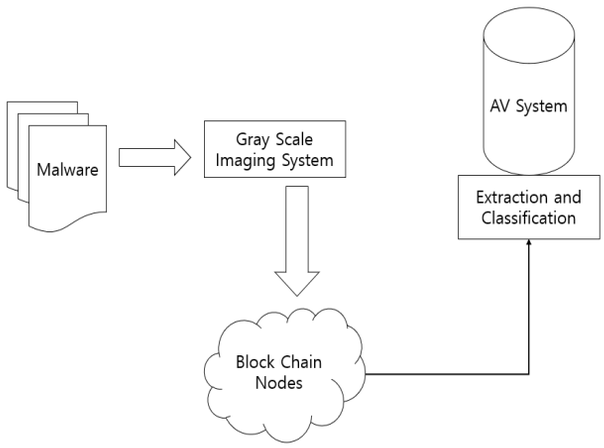

Fig. 9. Flowchart of AV System

transaction acceptance according to the transaction type and mutual consensus. In a consortium blockchain, the ledger is established according to the rules agreed by the consortium members. In terms of records management, a consortium blockchain, like a private blockchain, has an administrator subject. However, it is not a single entity but a group entity. As such, records are highly reliable, and permissions can be set and assigned to ensure the privacy of transaction records. Unlike private blockchains, it is difficult for an administrator of an organization to manipulate the specific record arbitrarily, increasing the reliability of records. However, the fact that a group of entities must have the authority and consent to the transaction and that the subjects must be on the same level in the hierarchy of the organization makes it difficult to be applied to the entire record management process. As such, it is necessary to find ways to converge it or use it in parallel with other types of blockchains.[10]

2.7 Domestic and Overseas Trend of Blockchain

The blockchain that is based on cryptocurrency is most actively discussed in the financial industry, but it has recently been accepted as the innovative technology to supplement the shortcomings of non-financial and financial industries. Furthermore, the blockchain is converged with fintech and other technologies to be applied to various industries, and it has been promoting and expanding the application and changes in various industries through the integration of innovative ideas and technologies by

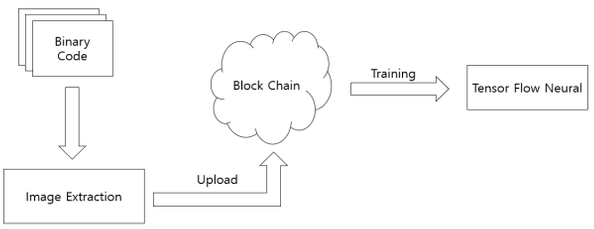

Fig. 10. Extraction Flowchart

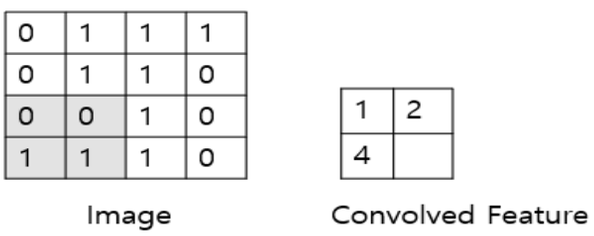

Fig. 11. Image Convolve

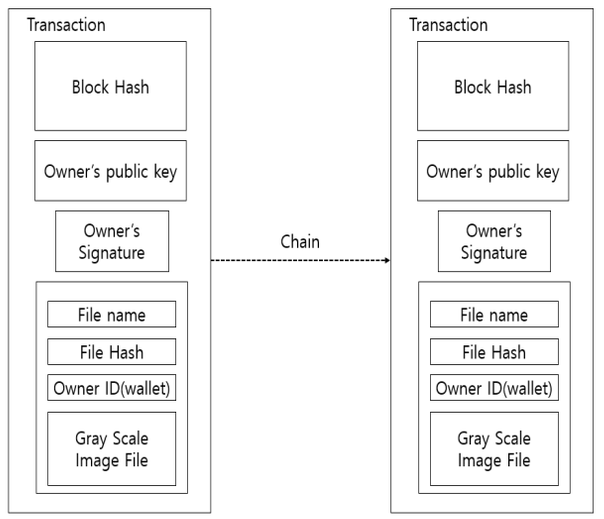

Fig. 12. Transaction Block

startups and cooperation led by financial industry and government.[11]

3. Proposed System

This Fig. 9 shows the flow of the process of the system proposed in this paper and how it classifies and learns the AV (Anti-Virus) system. Firstly, it registers the image in the blockchain. Users convert files into images with the module that extracts the grayscale of the file and register them in the blockchain, and the blockchain retrieves the registered images and classify them as malware or normal file. The method proposed in this paper uses the blockchain technology to add more data in real-time, enables more advanced machine learning through a wide range of datasets, and assures the diversity of the exhaustive search of the conventional deep learning method.

Fig. 10 shows the tensor flow neural network and the flow of the classification system that converts the malware into the image. Firstly, it converts the byte code of malware into an image. The image has a specific size, which affects the accuracy of the classification. The system then registers the image of the malware in the blockchain and learns the extracted malware image with the tensor flow neural network. Using the blockchain to learn the classified files and pattern can increase the accuracy of the future classification.

The detection step determines whether the file inputted by the learned model is a normal file or malware. It checks the grayscale image converted from the byte code of malware. In the case of many malware and variant families, images in the same family show very similar results in layout and texture. The system extracts characteristics through the image convolve and determines similarities.

This paper proposes the blockchain to replace the DB server in existing AV systems. Existing DB servers are not open to users who use the software, and it can be easily exposed to malware unless the AV developer provides the latest update before the DB server examines the sample file. If blockchain replaces the DB server, it can build the samples in real-time and can enhance the security. Moreover, it can conduct the exhaustive search widely and quickly compared to other deep learning systems and can learn from more diverse situations. Fig. 12 shows the structure of blocks registered in the blockchain. It has the basic structure of blockchain, and the filename, file hash value, registrant, and image are registered by the part that uploads the image. Since a blockchain can experience the deterioration of storage space and speed as the block size increases, uploading grayscale images can use the storage space more efficiently than uploading files.

Since the security of the blockchain in the system proposed by this paper can increase when there are

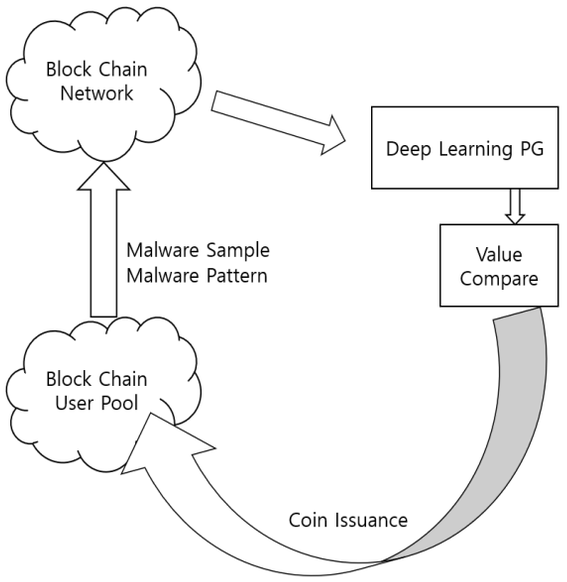

Fig. 13. Flowchart of Blockchain Contract

Table 1 Comparison with Existing Systems

	DB server AV	Block Chain Deep Learning AV	Deep Learning AV
Security	△	○	○
Diversity	×	○	×
Real-time cast	×	○	×
Speed	○	×	△

more participants, the system needs many users. As such, the system may offer incentives to users, who registered the blocks, to motivate them to participate more as a way to promote the blockchain. The blockchain system proposed in this paper uses a smart contract to encourage participation. When a user registers a block in the blockchain, the deep learning system extracts the value of the registered malware or pattern and excludes the existing malware to solve the duplicated registration. The system can pay the coin to the users who register the unregistered new malware to increase user participation. It can disclose the code related to reward payment using the smart contract to improve reliability.

Table 1 shows the comparison of the conventional DB server-based AV system, deep learning AV system without the blockchain, and the AV system proposed in this paper. The existing DB server-based AV system is vulnerable to security since it can be attacked by the latest or variant malware if it fails to continuously update the malware signature and other data to upload to the DB server. Moreover, it is not a real-time system since it analyzes and classifies sample files before registering them in the DB, and it takes time to update. However, it is excellent in speed compared to deep learning systems since it registers the new malware information in the DB and compares the file with the registered DB without requiring the time for learning. The next is the malware analysis and detection systems using deep learning. Unlike other systems, the deep learning system features excellent security since it detects new malware and variant malware through machine learning. However, its performance is weak from the diversity and real-time aspect since it must examine the samples needed for deep learning exhaustively. Its speed is not as good as the existing AV systems. A system using the DB is much superior to deep learning systems in speed since it uses the method of detecting malware by comparing it with the registered database. It is faster than the blockchain AV system proposed in this paper also. However, the blockchain system is much superior in the security aspect since it classifies, learns, and detects the exhaustively examined malware through deep learning. Unlike the conventional exhaustive search method, it can identify a wide range of malware types since the malware can be registered from anywhere in the world. Moreover, it uses the real-time nature of blockchain to register and learn malware faster than other systems. Since the diversity of samples increases as the users of the blockchain increases, the diversity of malware leaning is also excellent. However, the blockchain is slow, and the learning time increases as time passes since the length of the blockchain increases geometrically. Therefore, the overall speed decreases. Although its speed is weaker than other systems, it features superior security, diversity of malware detection, and real-time aspect with the blockchain than other systems. This paper proposes the blockchain technology to supplement such weaknesses of other systems.

3.1 Sustainability

[18] presented a survey of technologies implemented in smart cities, including edge computing, Blockchain,

and AI. These smart technologies help establish a sustainable environment by reducing bandwidth usage, latency, and power consumption in Internet of Things (IoT) based devices operating various smart city-based applications. The method proposed in this paper helps to build a more sustainable environment than the existing one by saving the manpower and time needed to define the existing anti-virus as a signature or act. Blockchain can be used to reduce the manpower and power required to collect data, while maintenance costs can also be significantly reduced compared to traditional methods, helping the environment in which the system is built and operated. It also has more economical effects than developing new systems and signatures using deep learning.

In other words, the ongoing operation of the system was studied using blockchain and deep running in order to compensate for the limitations of the past systems with the appearance of various malicious codes and management personnel as problems with the continuous operation of the existing system. Thus, the design of the system in this paper can be viewed positively with respect to the ongoing utilization of computing, and can also be expected to perform better.

4. Conclusion

Damages by malware have recently been increasing. Conventional signature-based antivirus solutions are helplessly vulnerable to unprecedented new threats such as Zero-day attack and ransomware. This paper proposes a solution using the blockchain and deep learning technologies as the next-generation antivirus solution. It supplements the conventional AV systems that compare and detects malware using the DB server and AV software by replacing the DB with the blockchain. It uses deep learning of various samples and forms to increase the detection rate of continuously new and variant malware. Moreover, it can supplement the existing deep learning method requiring various learning and exhaustive search with the blockchain and can link with the smart contract to continuous encouragement more participation by users. It shows the direction for the development of next-generation AV solutions.

References

[1] Why even the best antivirus software isn't enough (and why you still need it) [Internet], https://www.csoonline.com/article/3316480/why-the-best-antivirus-software-isnt-enough.html

[2] G. S. Kang, "Study on Cloud Computing-Based Malware Detection System," Master, Konkuk University, Korea, 2015.

[3] S. Y. Choi, "Emulation-Based Abnormal Web Page Link Analysis to Detect Malware-istributing Networks," Master, Jeonnam National University, Korea, 2014.

[4] Symantech - Internet Security Thresat Report [Internet], https://docs.broadcom.com/docs/istr-24-executive-summary-en

[5] A. Souri and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques," *Humand-centric Computing and Information Sciences*, Vol.8, No.3, 2018.

[6] A. Douglas, R. Holloway, J. Lohr, E. Morgan, and K. Harfoush, "Blockchains for constrained edge devices," *Blockchain: Research and Applications,* Vol.1, Iss.1-2, 2020.

[7] S. W. Kim, "Real-Time Malware Detection in Intrusion Detection System," Master, Hanyang University, Korea, 2014.

[8] J. W. Chang, "Study on Detecting Malware and Security Control Measures," Master, Korea University, 2014.

[9] J. H. Park, "Effect of Private Blockchain Characteristics on Acceptance in Medical Sector," Master, Graduate School of Sungkyunkwan University, Korea, 2018.

[10] G. Y. Lee, "Study of Applying Blockchain Technology and Record Management System," Master, Myongji University, Korea, 2019.

[11] P. S. Goh, "Study on Use of Medical Information System with Blockchain," Master, Soongsil University, Korea, 2019.

[12] D. G. Guh, "Study on Predicting Taxi Passenger Using Deep Learning," Master, University of Seoul, Korea, 2018.

[13] D. T. Ramotsoela, G. P. Hancke, and A. M. Abu-Mahfouz, "Attack detection in water distribution systems using machine learning," *Human-centric Computing and Information Sciences*, Vol.9, No.13, 2019.

[14] S. D. You, C. H. Liu, and W. K. Chen,"Comparative study of singing voice detection based on deep neural networks and ensemble learning," *Human-centric Computing and Information Sciences*, Vol.8, No.34, 2018.

[15] S. N. Danilin and S. A. Shchanikov, "Neural network algorithms for determining the values of signal parameters in radio- electronic hardware," 2017 Dynamics of Systems, Mechanisms and Machines (Dynamics), Omsk, 2017, pp.1-4, 2017.

[16] Y. B. Cho, "Detection Technique of Malware Using Deep Learning-Based R-CNN," Master, Daejeon University, Korea, 2018.

[17] Y. Cheong, "Blockchain-Based Image Information Management System," Master, Ajou University, Korea, 2019.

[18] S. I. Jung and H. W. Kim, "Web-Anti-MalWare Malware Detection System," *Proceedings of the Korean Society of Computer Information Conference*, pp.365-367, 2014.

[19] D. Lee and J. H. Park, "Future Trends of AI-Based Smart Systems and Services: Challenges, Opportunities, and Solutions," *Journal of Information Processing Systems*, Vol.15, No.4, pp.717-723, 2019.

Deok Gyu Lee

https://orcid.org/0000-0003-4057-9558

e-mail : deokgyulee@gmail.com

He received the M.S and Ph.D. degree in Graduate School of Computer Science from SoonchunhyangUniversity, Korea in 2001 and 2006 respectively. From 2006 to 2014 he worked as a senior member of engineering at the-cyber security research division in the ETRI(Electronics and Telecommunication Research Institute), Korea. Since 2014,he is currently an associate professor in the Department of Information Security, Seowon University. Dr. Lee haspublished many research papers in international journals and conferences. Dr. Lee has served as Chairs, program committee for many international conferences and workshops.