JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 4, pp. 9-17, April 2022
https://doi.org/10.9708/jksci.2022.27.04.009

A Conceptual Architecture for Ethic-Friendly AI

Yustus-Eko Oktian*, Stanley Brian**, Sang-Gon Lee***

*Post Doc. Researcher, Blockchain Platform Research Center, Pusan National University, Busan, Korea
**Students, Dept. of Computer Science of Graduate School, Dongseo University, Busan, Korea
***Professor, Dept. of Information Security, Dongseo University, Busan, Korea

[Abstract]

The state-of-the-art AI systems pose many ethical issues ranging from massive data collection to bias in algorithms. In response, this paper proposes a more ethic-friendly AI architecture by combining Federated Learning(FL) and Blockchain. We discuss the importance of each issues and provide requirements for an ethical AI system to show how our solutions can achieve more ethical paradigms. By committing to our design, adopters can perform AI services more ethically.

▸ **Key words**: AI, AI Ethics, Architecture, Blockchain, Federated Learning

[요 약]

최첨단 AI 시스템은 방대한 데이터 수집에서 알고리즘 편향에 이르기까지 많은 윤리적 문제를 드러내고 있다. 이에 본 논문에서는 연합학습과 블록체인을 결합하여, 더 윤리적인 AI 아키텍처를 제안하였다. AI의 윤리성에 관한 중요한 문제들을 논의하고, 문헌조사를 통하여 윤리적 AI 시스템에 대한 요구사항을 연구하고 도출한다. 제안한 아키텍처의 요구사항 만족을 분석하였다. 제안한 AI 구조를 디자인에 채택함으로써 AI 서비스를 보다 윤리적으로 수행할 수 있다.

▸ **주제어**: AI, AI 윤리, 아키텍쳐, 블록체인, 연합학습

• First Author: Yustus-Eko Oktian, Corresponding Author: Sang-Gon Lee
*Yustus-Eko Oktian (yustus@islab.re.kr), Blockchain Platform Research Center, Pusan National University
**Stanley Brian (briansta1601@gmail.com), Dept. of Computer Science of Graduate School, Dongseo University
***Sang-Gon Lee (nok60@dongseo.ac.kr), Dept. of Information Security, Dongseo University

• Received: 2021. 12. 30, Revised: 2022. 04. 19, Accepted: 2022. 04. 19.
• This paper is an extended version of the paper presented at the "21st Information Security Conference" hosted by the Korean Institute of Information Security & Cryptology in February 2021.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

We long time believe that users only have constrained machine that is unable to train the machine-learning model. They also have a limited data size, which is not enough to produce a highly accurate training model. Therefore, many data from different users must be aggregated to a high-performance server, where the training takes place. In consequence, users lose control of their data once the data is transferred out from their devices. This data collection practice sometimes does not explicitly request user consent. Many companies use an “opt-out” mechanism instead of “opt-in”, which puts users on surprise when they realize such a data collection setting exists. Even worse, there is no regulation for companies when conduct AI practices. Until recently, the public became aware of the importance of user privacy with the introduction of the GDPR law [1].

Even though massive data collection can be compelling, it is challenging to adjust the trade-off between the benefits and user privacy. For example, China’s social credit system [2] can shape the business and citizens’ behavior towards better goals (in the view of the government). However, the citizen is at a disadvantage by losing freedom over this mass surveillance program. Moreover, AI is a black box system (in the current form), making it very tough to be debugged. This problem leads to many biases in AI algorithms. For instance, South Korea AI persona, Lee Luda [3], makes a controversy because she used offensive language targeting a minority community. Amazon AI recruitment tools are also being shut down because it prefers men over women in selecting candidates [4]. As a result, researchers and AI practitioners must conduct AI services with ethics-in-mind, which always preserve human values.

This paper aims to seek solutions towards more ethic-friendly AI architecture by combining Federated Learning (FL) [5] and Blockchain [6]. FL preserves user privacy by training private user data on user local machines instead of sending them to the server. Meanwhile, the blockchain serves as a trusted platform to conduct the overall FL process so that FL participants can collaborate in a secure, transparent, and fair manner. We also aim to analyze the importance of issues in AI ethics, identify the requirements of an ethical AI system and finally show that our solutions tackle the necessary components. By committing to our design, adopters can realize an ethic-friendly AI architecture.

II. Important Issues in AI Ethics

Hagendorff et al. [12] in his paper ranked the number of iterations of researched issues in AI ethics among research papers, which resulted in the top 5 rankings in order as listed in Table 1.

Table 1. Summary of issues and explanations in AI ethics.

Rank	Issue	Explanation
1	Privacy protection	Importance of ethics in the handling of the user’s data
2	Fairness	Ensure individuals and groups are free from unfair bias on decisions
3	Accountability	Responsible for the decision making of created AI systems
4	Transparency/ openness	Openness to the public analysis of the algorithms, data, and design processes of the AI system
5	Safety/ cybersecurity	Secure and resilient to intelligent adversaries

The most popular talking points of privacy issues in AI mention the importance of ethics in the handling of the user’s data. If not regulated properly, collected data can be sold to third-party members and utilized to track their behavior for personal gain [13]. Transparency in the privacy policy is also one of the important aspects of AI systems implementation. Those who implement AI need to demonstrate how their AI will use and process data, allowing the customers to make decisions to determine if their approach is ethical or not to their circumstances. Thus, privacy corresponds with transparency.

High-Level Expert Group on Artificial Intelligence (HLEG), which is an independent expert group that was set up by the European Commission, addresses fairness as the issue of AI playing a part in ensuring individuals and groups are free from unfair bias, discrimination, and stigmatization [9]. Additionally, this coincides with the issue of accountability, that AI developers are required to take responsibility for the fairness of the decision-making system that was developed. Addressing the issue of accountability, HLEG also mentioned the auditability, meaning for the openness to analyzing the algorithms, data and design processes of the AI system, solving the transparency/openness issue.

On the topic of cybersecurity, HLEG added that AI practitioners also need to question whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behavior of their AI systems. Holdren et al. [14] mentioned that those involved in AI issues should engage with their cybersecurity colleagues for input on how to ensure that AI systems and ecosystems are secure and resilient to intelligent adversaries and collaborate to introduce innovative ways to apply AI for effective and efficient cybersecurity. In conclusion, solutions toward AI ethics are recommended to gravitate toward the essence of privacy, accountability, cybersecurity based on how they universally cover other issues.

III. Proposed Architecture

We propose an ethic-friendly AI architecture as depicted in Fig 1. The proposed system comprises six components: model owners (e.g., AI companies), local trainers (e.g., users), verifiers (e.g., users or government personnel), trusted auditors (e.g., the government), peer-to-peer (P2P) blockchain, and P2P storage. All participants are authenticated and endorse the use of a reputation system in our system. The AI workflow is described as follows.

Registration: The government makes the digital representation of the training policy in smart contracts (Step 1). AI companies, as model owners, create an initial global model and prepare rewards for trainers. They then create a training task in the smart contract (Step 2). After that, the companies request approvals from the government (Step 3). Before approving a task, the government must make sure that the proposal provides enough incentives for trainers. They also create a standardized test dataset suitable for the proposal (Step 4). The model parameters and the test dataset will be distributed to trainers and verifiers through the P2P storage. Meanwhile, the hash of the model and dataset is stored in the blockchain.

Training: Users can join the training as trainers by registering themselves in the smart contract (Step 5). They can then get the model from P2P storage (Step 6) and begin training using their local data (Step 7). When the training is complete, users submit the trained model through P2P storage while the hash is logged in the blockchain (Step 8).

Evaluation: Users or government personnel can register themselves as verifiers in the smart contract (Step 9). At each global epoch, the verifiers must get the test dataset (Step 10) and the trained local models (Step 11) from P2P storage. They then verify the accuracy of the trained models using the test dataset (Step 12). Once the evaluation finishes, the evaluation result is submitted to the smart contract (Step 13).

Aggregation: When a particular global epoch finishes, the companies get all of the trained local models from the P2P storage (Step 14). They then retrieve all of the associated evaluation scores from the smart contract (Step 15). Using the evaluation scores as a guideline, the companies aggregate the models according to their

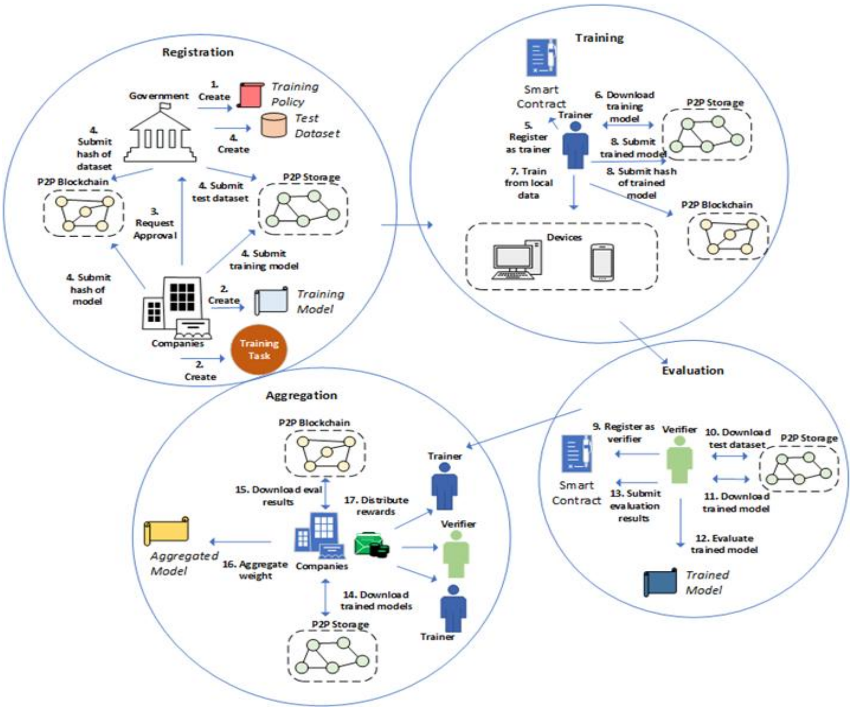


Fig. 1. The proposed ethic friendly AI system.

contributions (Step 16). For example, they may skip models with low accuracy as they are most probably trained with poisoned data or low-quality data. During evaluations, verifiers use adversarial defense techniques to check if the model is trained with adversarial examples. Therefore, the companies must also skip models, which contains malicious flag from the verifiers. Once the aggregation is completed, the companies distribute the reward to all trainers and verifiers through the smart contract (Step 17).

IV. Ethic-Friendliness Analysis

A. Requirements for Ethical AI

Floridi and Taddeo [7] divides ethics of AI into three spheres: ethics of data, ethics of algorithms, and ethics of practices. The explanation of each points are detailed in Table 2.

Ethics of Data: The ethics of data focuses on the ethical problems related to data, including generation, curation, processing, dissemination, sharing, and usage [7]. Tranberg et al. [8] recommends five principles to enforce data ethics: R1) human being at the center, R2) individual data control, R3) transparency, R4) accountability, and R5) equality.

Ethics of Algorithms: The ethics of algorithms addresses issues posed by the increasing complexity and autonomy of the AI algorithms [7]. HLEG mentioned that AI algorithm must follow these ethical principles [9]: R6) respect for human autonomy, R7) prevention of harm, R8) fairness, and R9) explicability.

Ethics of Practices: The ethics of practice focuses on the pressing questions about the responsibilities and liabilities of people and organizations in charge of data, strategies, and policies of AI systems [7]. Google provides a recommended practice for AI [10], which includes: R10) use a human-centered design approach, R11) rigorous testing, and R12) continuous monitoring and updates.

From these 3 groups of ethics, we propose a federated learning schemes as a solution for the

Table 2. Requirements for ethical AI

Requirement No	Issues	Explanation
R1	Human being at the center	The ethics of AI stsyem must adhere to human interest
R2	Individual data control,	A person should be in control of how his/her data will be processed
R3	Transparency,	Processing of data must make sense to the individual owning the data
R4	Accountability	The party in charge of processing the data is responsible toward the consequences of the data usage
R5	Equality	Extra care have to be made during the data process, to make sure individuals vulnerable to discrimination based on the results are are unharmed.
R6	Respect for autonomy	AI systems should complement human activities not instead controlling and sreering them.
R7	Prevention of harm	AI systems should be designed to prioritize human safety.
R8	Fairness	AI system shluld provide equal opportunity and benmefits not discriminationg based on race and gender
R9	Explicability.	AI system must make effort on making their process to be sufficiently understood
R10	Human-centered design approach	Processes need to be transparent
R11	Rigorous testing	Systems mustbe tested properly before launch to ensure of unnecessary bugs.
R12	Continuous monitoring and updates.	Systems are maintained and continuously patched for vulnerabilities

privacy issues corresponding to each points.

In conclusion, Ethics of Data corresponds to privacy and transparency of data handling, while Ethics of Algorithms and Practices details the importance of safety in AI.

B. Ethics Analysis of the Proposed Solutions

The connection between the proposed solutions and ethics requirments are detailed in Table III.

Training distributedly using Federated Learning: Users train their data locally on their devices and only send the model parameters instead of the private data to the server. The server then combines the trained local models into a single global model using an aggregation algorithm (e.g., Federated Averaging [5]). Using this approach, the user data do not leave the devices, and users still have control over their data (i.e., solving R2).

Rigorous evaluation and auditing: To ensure the quality of the trained models, they must be evaluated. For this purpose, we employ the government and volunteers as our verifiers. The government must first create a standardized training policy for AI companies in the form of federal or international law (e.g., GDPR [1]). With this law, we can hold malicious persons or organizations accountable (i.e., solving R4). We can also ensure that the AI models will always benefit humans (i.e., solving R1, R6, and R10). Moreover, the government must produce a generalized test dataset to be used during the evaluation stage. Assuming that this standardized test dataset has a high variance to cope with all possible classes, then this test should mitigate the AI bias that may happen during training (i.e., solving R5 and R8).

The group of verifiers evaluate the submitted local models from users to detect potential poisoning attacks on each epoch. Attackers can intentionally train the local model with bad or low-quality data to reduce the global model's overall accuracy. Moreover, the attackers can also train the model with adversarial examples to make the global model misclassify particular targets. Once detected, the attackers will be punished economically or by law (i.e, solving R7, R11, R12).

Logging training processes using the blockchain: In our architecture, all of the training processes are logged in the blockchain (e.g., Ethereum [6]).

Table 3. Connection between requirements and the solutions provided

Requirements	Solutions
Ethics of Data R1) Human being at the center R2) Individual data control, R3) Transparency, R4) Accountability, and R5) Equality	Users train their data locally on their devices and only send the model parameters instead of the private data to the server (R1, R2)
	All data in the blockchain is open to all the blockchain nodes, allowing auditors to examine the federated learning processes (R3)
	With a standardized training policy created by the government we can hold attackers accountable (R4)
	Assuming that this standardized test dataset has a high variance to cope with all possible classes, then this test should mitigate the AI bias that may happen during training (R5)
Ethics of Algorithms R6) Respect for autonomy, R7) Prevention of harm, R8) Fairness R9) Explicability.	With a standardized training policy created by the government we can hold attackers accountable, preventing harm caused by corrupted models (R6, R7).
	Assuming that this standardized test dataset has a high variance to cope with all possible classes, then this test should mitigate the AI bias that may happen during training (R8)
	All data in the blockchain is open for all the blockchain nodes, allowing auditors to examine the federated learning processes (R9)
Ethics of Practices R10) Human-centered design approach R11) Rigorous testing R12) Continuous monitoring and updates.	With a standardized training policy created by the human-centered government we can hold attackers accountable, design approach ensuring the AI models will benefit humans (R10).
	Malicious attackers can be detected based on verifiers' evaluation and will be punished by law (R11, R12).

Because of the chain-of-hashes introduced in the blockchain, the stored data in the blockchain becomes hard-to-tamper. All nodes must also include their digital transactions when storing data in the blockchain. Hence, malicious entities can be detected easily. Finally, all data in the blockchain is open for all the blockchain nodes. Hence, solving R3 and R9.

C. Comparison with Traditional FL Schemes

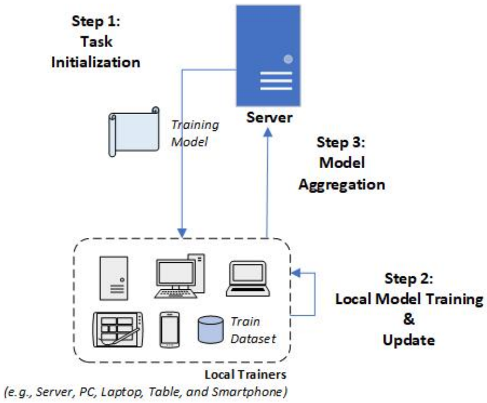


Fig. 2. Traditional FL scheme

In this section we compare our proposed federated learning scheme with the commonly used federated learning schemes [21]. A common FL scheme starts by setting up multiple clients to train a model by utilizing their owned datasets monitored by a centralized server. Clients send update parameters instead of raw data to the FL server. As a result, FL systems have substantially lower latency than centralized systems. FL clients train a global model utilizing their joined data without disclosing each device's personal information to the centralized server. The three-step approach shown in Fig 2 is utilized to achieve this scheme.

Task initialization: In a specific interval, the server selects a certain number of devices from the thousands available. It determines the target application and data requirements once the training task is specified. In addition, the server sets the hyperparameters related to the model and training process, such as the learning rate. Specifically, it initializes the weights on the server by leveraging weight initialization methods.

Local model training: Each participants receive the global model, and updates the local model parameters by training the model based on their device data. The objective of the clients are therefore to obtain an optimal parameter at current time iteration. Finally, each local model's updated parameters are sent again back to the FL parameter server.

Model aggregation: The centralized server receives the local parameters from each participant and aggregates the local models from the participants, then sends the updated global model parameters back to all the participating clients to minimize the global loss function.

Compared to this scheme, our proposed scheme in Fig 1. added several mechanisms to improve the original scheme. First, the original scheme uses a centralized endpoint to aggregate the updated models to the server. This approach opens vulnerability to malicious attacks and is susceptible to single-point-of-failure problem. Our proposed scheme adds a blockchain and peer-to-peer storage to provide a decentralized aggregation method. Second, without an incentive to train, it will be difficult to find participants who are willing to lend their devices for model training. In our scheme, by involving the company to provide a reward policy and using the smart contract available through the blockchain (such as Ethereum) we can add an incentive for participants to join the model training. Lastly, we cannot detect and punish trainers in the original scheme for poisoning the models with malicious data.

Our scheme provides solution by allowing clients to join as a reviewer and verify the quality of trainers' contribution. Reviewer's work are rated depending on their precision, with the system punishing reviewers deemed malicious with their given scores are too high or too low, or rewarding reviewers with precise scores.

D. Feasibility of Proposed Scheme

The use of blockchain for privacy-preserving machine learning has been examined and researched in several works, among those are [15 ~ 20]. CrowdSFL [15] provide a framework for crowdsourced federated learning and im-plemented blockchain as a method of storing the model. The work also examined the efficiency of BCFL. FLChain[16] proposes a framework for performing Federated-Learning in Mobile Edge Computing with the aggregation performed by blockchain. Another work, DeepChain [17] came up with a Federated Deep Learning framework with blockchain as an incentive of work for clients. Jonathan et al. [18] proposed a blockchain system to improve the security of a federated learning scheme used in healthcare. Nguyen Quang et al. [19] proposed a framework about managing the resource-usage in blockchain-enabled federated learning. Y. Lu et al. [20] came up with for including blockchain to improve an existing FL framework running in digital twin wireless networks (DTWN) for collaborative computing, which improves the overall reliability and security of the system and privacy. With the demonstrated amount of variety and scope of works in blockchain-enabled federated learning, it can be concluded that blockchain is a viable approach for improving the process of federated learning.

V. Conclusions

This paper proposed a more ethical AI architecture through a combination of federated learning and blockchain technologies. The federated learning yielded promising solutions towards AI ethics in terms of data collection and training transparency. Meanwhile, the blockchain enhanced AI ethics with its secure, transparent, and fair collaborative auditing platform. Furthermore, with the addition of the role of verifiers as the evaluators of the trained model, malicious activities will be easier to be detected and punished accordingly, thus improving the overall security of the architecture. However, our proposal still does not solve AI's fundamental issues regarding its “black box” properties. More research towards “explainable AI” is still required in the future so that we as humans and AI supervisors can make a better decision on how to use AI.

ACKNOWLEDGEMENT

This work was supported by Dongseo University, “Dongseo Cluster Project” Research Fund of 2021(DSU-20210001).

REFERENCES

[1] M. Goddard, “The EU general data protection regulation (gdpr): European regulation that has a global impact,” International Journal of Market Research, vol. 59, no. 6, pp. 703–705, Nov. 2017. DOI: 10.2501/IJMR-2017-050

[2] K. Munro. (2018) China’s social credit system could interfere in other nations’ sovereignty. [Online]. Available: https://bit.ly/3qSmFz8 [Accessed: 25-Nov-2021].

[3] J. McCurry. (2021) South Korean AI chatbot pulled from facebook after hate speech towards minorities. [Online]. Available: https://bit.ly/2NwoZgM [Accessed: 25-Nov-2021].

[4] Reuters. (2018) Amazon ditched ai recruiting tool that favored men for technical jobs. [Online]. Available: https://bit.ly/3sTehBc [Accessed: 25-Nov-2021].

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR vol. 54, pp. 1273–1282, 2017.

[6] G. Wood et al., “Ethereum: A secure decentralised generalised transaction ledger,” Ethereum project yellow paper, vol. 151, no. 2014, pp. 1–32, 2014.

[7] L. Floridi and M. Taddeo, “What is data ethics?” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2083, p. 20160360, Dec, 2016. DOI: 10.1098/rsta.2016.0360

[8] P. Tranberg, G. Hasselbalch, B. K. Olsen, and C. S. Byrne. (2018) “Dataethics – principles and guidelines for companies, authorities, and organisations,” [Online]. Available: https://bit.ly/3canj6V [Accessed: 25-Nov-2021].

[9] AI HLEG. (2019) Ethics guidelines for trustworthy AI. [Online]. Available: https://bit.ly/3iTIRGs [AAccessed: 25-Nov-2021].

[10] Google. (2020) Responsible AI practices. [Online]. Available: https://bit.ly/3opdd4Q [Accessed: 25-Nov-2021].

[11] J. Benet, “Ipfs-content addressed, versioned, p2p file system,” arXiv preprint arXiv:1407.3561, July 2014.

[12] T. Hagendorff, “The Ethics of AI Ethics: An Evaluation of Guidelines”. [Online] Available: https://bit.ly/3l8jVgF [Accessed: 25-Nov-2021].

[13] University of Pennsylvania(2019) “Your Data Is Shared and Sold…What’s Being Done About It?”. [Online] Available: https://whr.tn/3HUbTBO [Accessed: 25-Nov-2021].

[14] Holdren, J. P., Bruce, A., Felten, E., Lyons, T., & Garris, M. (2016). “Preparing for the future of artificial intelligence,” pp. 1~58, Executive Office of the President National Science and Technology Council, Washington, D.C. 2016. Available: https://obamawhitehouse.archives.gov/blog/2016/05/03/preparing-future-artificial-intelligence [Accessed: 25-Nov-2021].

[15] Li Z, Liu J, Hao J, Wang H, Xian M., “CrowdSFL: A Secure Crowd Computing Framework Based on Blockchain and Federated Learning.” Electronics. vol. 9, no. 5, May 2020; p.773.

[16] Umer M. and Choong S. H. “FLchain: Federated Learning via MEC-enabled Blockchain Network.” the 20th Asia-Pacific Network Operations and Management Symposium: Management in a Cyber-Physical World, APNOMS 2019, pp. 1~4, 2019

[17] Jiasi W., Jian W., Jilian Z., Ming L, Yue Z., and Weiqi L., “DeepChain: Auditable and Privacy-Preserving Deep Learning with Blockchain-based Incentive,” IEEE Transactions on Dependable and Secure Computing, vol. 14, no. 8, p.1, Nov. 2019.

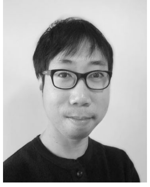
[18] Jonathan P. P., Tyler F., Robert M., Marielle S G., Heather L. F., and Bill G.. “A blockchain orches-trated federated learning architecture for healthcare consortia.” arXiv preprint arXiv:1910.12603, Oct. 2019.

[19] Nguyen Q. H., Tran T. A., Nguyen C. L., Dusit N., Kim D. I., and Erik E., “Resource Management for Blockchain enabled Federated Learning: A Deep Reinforcement Learning Approach.” Hieu, Nguyen Quang, Tran The Anh, Nguyen Cong Luong, Dusit Tao Niyato, Dong In Kim and Erik Elmroth. “Resource Management for Blockchain-enabled Federated Learning: A Deep Reinforcement Learning Approach.” ArXiv, ArXiv:abs/2004.04104 Apirl 2020.

[20] Y. L. Lu, X. H. Huang, K. Zhang, S. Maharjan, and Y. Zhang. “Low-latency Federated Learning and Blockchain for Edge As-sociation in Digital Twin empowered 6G Networks,” IEEE Transactions on Industrial Informatics, vol. 17, no.7 pp. 5098~5107, July, 2020, doi:10.1109/TII.2020.3017668.

[21] Abreha, H. G., Mohammad H., and Mohamed A. S.. "Federated Learning in Edge Computing: A Systematic Survey," Sensors, vol. 22, no. 2, p.450, Jan. 2022, https://doi.org/10.3390/s22020450

Authors



Yustus-Eko Oktian received his bachelor's degree in Electrical Engineering from Petra Christian University, Indonesia, in 2013, and his master's and doctoral degrees in Computer Engineering from Dongseo University, South Korea, in 2016 and 2021. Dr.Oktian is currently a post-doctoral researcher at Pusan National University, South Korea. His research interests are Network Security, Distributed Computing, Blockchain, Internet-of-Things (IoT), and Software-Defined Networking (SDN).

Stanley Brian received his bachelor degree in Informatics Engineering from Petra Christian University in 2019, Surabaya, Indonesia. Mr. Brian is a Master student at Dongseo University, Busan, Korea. His research interests are on the topics of blockchain, information security, and web-based applications.

Sang-Gon Lee received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Kyungpook National University, Korea, in 1986, 1988 and 1993, respectively.

Dr. Lee joined the faculty of the Department of Information Security at Dongseo University, Busan, Korea, in 1997. He is currently a Professor in the Department of Information Security, Dongseo University. He is interested in Blockchain, AI security, and Software Defined Network.