American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Opinion

Cryptography for Trusted Artificial Intelligence in Medicine

Macq B* and Quisquater JJ

Catholic University of Louvain, Belgium

***Corresponding author:** Macq B, Catholic University of Louvain, Belgium

To Cite This Article: *Macq B, Quisquater JJ. Cryptography for Trusted Artificial Intelligence in Medicine. Am J Biomed Sci & Res. 2021 - 13(2). AJBSR.MS.ID.001858.* DOI: 10.34297/AJBSR.2021.13.001858.

Received: June 16, 2021; **Published:** June 18, 2021

Introduction

On 8 April 2019, the High-Level Expert Group on AI appointed by the European Union produced an important document [1] describing key principles behind the deployment of an ethical AI. This document relies on seven key principles:

I. Human agency and oversight, Including fundamental rights, human agency and human oversight

II. Technical robustness and safety, Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility

III. Privacy and data governance, Including respect for privacy, quality and integrity of data, and access to data

IV. Transparency, Including traceability, explainability and communication

V. Diversity, non-discrimination and fairness, Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

VI. Societal and environmental wellbeing, Including sustainability and environmental friendliness, social impact, society and democracy

VII. Accountability, Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

Among these seven basic principles, some are clearly requiring the specific deployment of cryptographic tools and security architectures:

I. Resilience to attack requires architectures including authentication and integrity of data, ciphering of communication, integrity of the AI models, and authentication of the parties involved in the training and the use of the models

II. Respect for privacy, includes new learning models, being either federated (the data are kept at the source, and the model is travelling to learn from data silo to data silo) or based on the use of Full Homomorphic Encryption which allows the models to be executed directly on encrypted data without any decryption phase.

III. Traceability may require the authentication of the models (origin of the model and its integrity) and traceability of the model and data themselves. Emerging techniques of watermarking are suited for such challenges.

IV. Finally, fairness and stakeholders participation can be achieved by the blockchain technology.

Classical Security Architectures for Trusted AI

The basic tools of cryptography are suited for securing communication of the different steps in the deployment of an IA system. We will focus here on the sub-area of machine learning. For such systems, a first step consists in harvesting data to learn the model. The communications between the source of data and the learning model have to be secured ad minima by using classical cryptographic tools. The following steps can constitute an example:

I. The learning model is broadcasting its public key through a certificate

II. Each data source is generating and providing a session key encrypted by the public key of the learning model

III. The data batches provided by the sources are encrypted by their specific session key

IV. Each data batch is hashed for integrity check by the learning model

V. Each batch hash is signed by the secret key of the data sources

VI. The computed model after learning can be itself hashed and signed for integrity and origin verification

Such an approach can rely on exchanges at the http level and the use of Secure Socket Layer (https) and is straightforward to implement. The use of classical Advanced Encryption Standard (AES) ciphering, Secure Hashing Algorithms (SHA) for hashing and Elliptic Curves public key cryptography in SSL allows reaching a high level of confidence in each of these six steps.

However, such an architecture does not provide additional features, which are demanded by the requirements for a Trustworthy AI described in section 1.

Federated Learning, Federated Byzantine Agreements and the TCLearn Model

In a classical security approach described above, all the stakeholders have to be trustworthy. If the party managing the learning model is not trusted, it is required to distribute the learning in a secure federated learning by which the model is learned by travelling across the data sources. This prevents putting privacy at the risk of a leakage, since only the model is travelling on the network while the data remains at their source.

Distributed, federated learning [2] has been suggested for multiple applications, including the medical field [3]. This approach facilitates cooperation through coalitions where each member keeps control of and responsibility for its own data (including accountability for privacy and consent of the data owners such as patients). Batches of data are processed iteratively to feed a shared model locally. Parameters generated at each step are then sent to the other organizations to be validated as an acceptable global iteration for adjusting the model parameters. Thus, the coalition partners will optimize a shared model jointly by dividing the learning set into batches corresponding to the blocks of data provided by the members of this coalition.

The naive use of a CNN in a distributed environment exposes it to a risk of corruption (whether intentional or not) during the training phase because of the lack of monitoring of the training increments and the difficulty of controlling the quality of the training datasets. The distributed learning could be monitored by a centralised certification authority that would be in charge of validating each iteration of the learning process. Alternatively, a blockchain could be used to store auditable records of each and every transaction on an immutable decentralized ledger. The blockchain approach would provide distributed learning with a more robust and equitable approach for the different stakeholders involved in the learning process, since all of them are involved in the certification process.

We have recently proposed a new architecture, named TCLearn (Trusted Coalitions for Learning), that solves these security issues [4]. It is based on a blockchain technology using Federated Byzantine Agreements (FBAs). FBAs were described in the 1980s by L. Lamport et al [5]. This algorithm achieves trusted agreement among a community of users provided 2/3 of the members of the community are trustworthy.

FBAs are used in recent crypto-moneys such as Algorand and Stellar. They have also been described as a potential tool for solving issues in distributed learning when some data providers have unreliable connections (Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent).

Security and cryptographic issues have been or will soon be solved for blockchain-based distributed learning. However trustworthy distributed learning still faces open issues.

FBA distributed learning works properly if and only if the data quality is guaranteed in each member of the coalition. Setting peer-reviewed quality control of the data across the coalition, which should evolve in time with the progresses in the quality of image acquisition, is thus a requirement. Moreover, the FBA distributed architecture is suitable for collecting experts' opinions of algorithms. It is also suitable for setting distributed reinforcement learning algorithms for the design of new planning procedures within the coalition.

We show in (Figure 1), a practical implementation of FBA for the validation of iteration steps of a learning model, as deployed in TCLearn.

After a majority vote, the candidate model is accepted or not.

Deploying Machine Learning Through Homomorphic Encryption

Classical cryptography is a technique for secure communication between multiple parties, where one party encrypts a message and sends it to another party, who then decrypts it. In machine learning on sensitive data, for example personal health data, this technique allows to guarantee privacy only if the parties are trustworthy and robust against attacks. If not, a data source which sends its data to the model for learning, even encrypted, must appear in clear in the AI algorithm.

The goal of homomorphic encryption is to make feasible the required computations on data by AI algorithms, directly on encrypted data. Very concretely, the data are encrypted by a public key and send to a model. The model itself is encrypted with the same

public key and if the encryption is homomorphic, the computation done by the encrypted model, will produce an encrypted result, which is equivalent to a result, which would have been encrypted by the public key. Only the data provider that owns the corresponding secret key will be able to decipher the result. The AI model can therefore appear as an online service, making predictions directly on encrypted data and providing directly encrypted results, and ensuring the highest level of privacy.

The concept of computing over encrypted data was first introduced as a "privacy transformation" by Rivest, Adleman, and Dertouzos in 1978 [6] and developed in Europe by Paillier [7] and Izabachene [8] among others. A scheme is called additively (or multiplicatively) homomorphic if

$[x] \oplus [y] = [x+y]$ and $[x] \otimes [y] = [x \cdot y]$

for addition and multiplication, respectively. The symbols $\oplus$ and $\otimes$ respectively denote the homomorphic addition and multiplication operations in the ciphertext space. In other words, if an encryption scheme is additively homomorphic, then encryption followed by homomorphic addition is equal to addition followed by encryption.

A partially homomorphic encryption (PHE) scheme can perform a single operation, such as addition or multiplication, over encrypted data an arbitrary number of times. Fully homomorphic encryption (FHE) enables arbitrary addition and multiplication over encrypted data.

An Example for the Screening of Patients Against Neurodegenerative Diseases

Recent advances in Full Homomorphic Encryption create new prospects for using personal data for screening and diagnosis. In FHE, personal data are encrypted through a public key associated with an individual. The data are processed by a model enciphered with the individual's public key. The result is obtained in an encrypted mode and is accessible only to the owner of the private key associated with the public key that was used. This makes it possible to constitute a personal homomorphic vault containing data over a long period of time. The encrypted data will be processed by models that could be improved over time through machine learning.

An example that is explored at UCLouvain on the early detection of Alzheimer's disease from personal data is given in (Figure 1). The model is trained "in clear" by data coming from volunteers, while the tests for large scale screening are achieved with a high level of privacy by using the FHE.

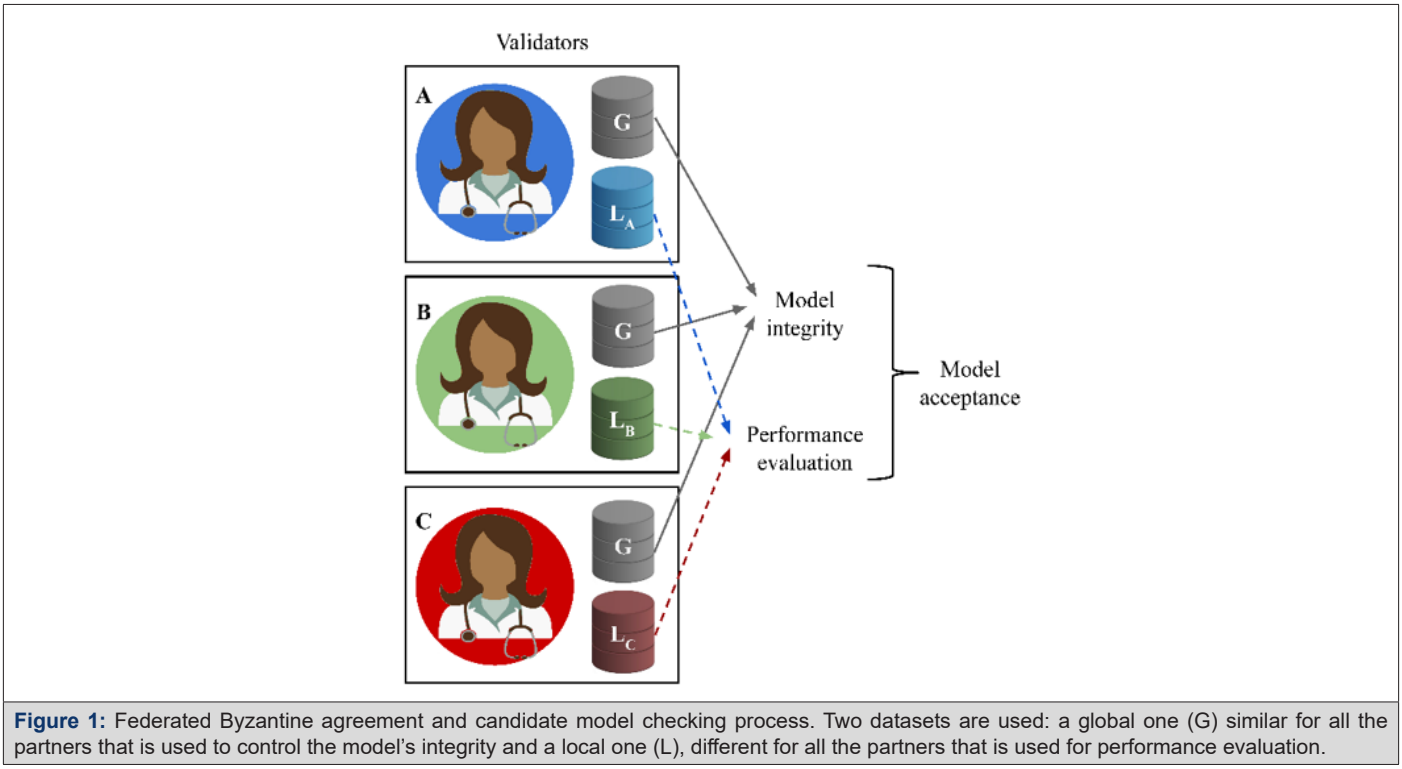

Figure 1: Federated Byzantine agreement and candidate model checking process. Two datasets are used: a global one (G) similar for all the partners that is used to control the model's integrity and a local one (L), different for all the partners that is used for performance evaluation.

Traceability of Data and Models Through Watermarking

The watermarking of data can be used for tracing their leakage. Each batch of data can be slightly but robustly modified to contain a specific mark of the destination (for example the learning model) in such a way that leakage at the destination point can be detected (Figure 2). Watermarks can also be used as seals to authenticate the data.

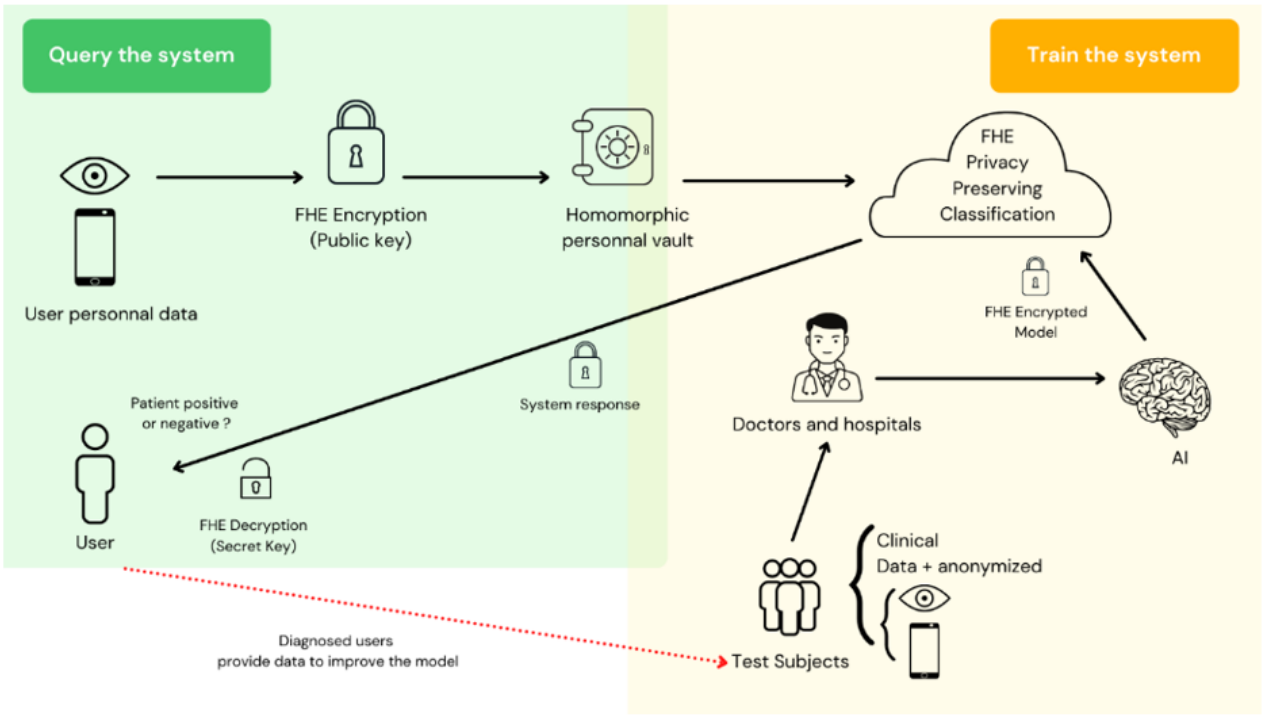

Figure 2: A privacy-preserving classification model. A user encrypts her private medical data and uploads it to some cloud computing service. Researchers encrypt their trained model parameters under the same public key and upload to the cloud, which performs the required computation for privacy-preserving classification. This result is then returned to the user, who decrypts it with his/her private key. Some of the patients agree to participate in clinical trials for which they give their complete data sets for AI training in an anonymized environment based on batching and federated learning.

Some recent works are also studying the possibility to insert watermarks in AI models: the weights of the neural networks are slightly modified to individualize each instance of the model. The watermark can be used to trace forgeries of the model or to be used as an authentication seal [9].

Conclusions

Cryptography for AI is a very active field which relies on emergent tools like full homomorphic encryption, watermarking of data and models and FBA-based blockchains. This research is mandatory to converge towards ethical trusted AI.

References

1. Ethics guidelines for trustworthy AI.
2. Li T, Sahu AK, Talwalkar A, and Smith V (2020) Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine 37(3): 50-60.
3. Rieke N, Hancox J, Li W, Milletari F, Holger R, et al. (2020) The future of digital health with federated learning. NPJ Digit Med 14(3); 1-7
4. Lugan S, Desbordes P, Brion E, Tormo LXR, Legay A, et al. (2019) Secure architectures implementing trusted coalitions for blockchained distributed learning (TCLearn). IEEE Access 7; 181789-181799.
5. Lamport L (1983) The weak Byzantine generals problem. Journal of the ACM (JACM) 30(3); 668-676.
6. Rivest RL, Adleman L, and Dertouzos ML (1978) On data banks and privacy homomorphisms. Foundations of secure computation 4(11); 169-180.
7. Paillier P (1999) Public-key cryptosystems based on composite degree residuosity classes. In International conference on the theory and applications of cryptographic techniques. Springer Berlin Heidelberg 223-238.
8. Chillotti I, Gama N, Georgieva M, and Izabachene M (2016) Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds. In international conference on the theory and application of cryptology and information security. Springer Berlin Heidelberg 3-33.
9. Uchida Y, Nagai Y, Sakazawa S, and Satoh SI (2017) Embedding watermarks into deep neural networks. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval 269-277.