|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATION****STANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-B-016 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | New York, 15-16 Nov 2018 |
| **DOCUMENT** |
| **Source:** | Lausanne University Hospital (CHUV), EPFL |
| **Title:** | Proposal: Secure and privacy-preserving benchmarking for artificial intelligence in health |
| **Purpose:** | Discussion |
| **Contact:** | Jean Louis RaisaroLausanne University Hospital (CHUV)Switzerland | Tel: +41 79 556 73 65Email: jean.raisaro@chuv.ch |
| **Contact:** | Jean-Pierre HubauxEPFLSwitzerland | Tel: +41-21-693-2627Email: jean-pierre.hubaux@epfl.ch |

|  |  |
| --- | --- |
| **Abstract:** | Artificial intelligence has the potential to revolutionize the field of healthcare in the next few years. Yet, its success highly depends on how AI algorithms for health are going to be benchmarked and regulated and on the trust that people have that this process is safe and secure. In this short document, we outline a proposal for a privacy-preserving benchmarking pipeline that leverages advanced privacy-enhancing technologies such as homomorphic encryption and distributed ledgers. |

# 1. Introduction

With the massive digitalization of medical data and the impressive advancements in artificial intelligence, it is becoming evident that AI-based algorithms will play an increasingly important role in the next-generation decision support systems.  Yet, to fully realize this potential, the health community has to face several challenges that range from the interpretability and explainability of models to the privacy and security concerns that stem from the sensitive nature of the data used to train and test these models.

As described in its first white paper [1], the Focus Group on Artificial Intelligence for Health (FG-AI4H) is working towards a standardized assessment of AI-based solutions for health, which will assure its quality, foster the adoption and have a strong impact on healthcare at the global level.  For this purpose, a tentative benchmarking pipeline that will be applicable to many different scenarios has been proposed. This pipeline consists of several steps such as (i) model training on a combination of private (owned by the researcher) and public data, (ii) submission of trained models to a benchmarking platform, (iii) evaluation of models on undisclosed test data, and (iv) publication of results on central leader board. The credibility of the benchmarking process highly depends on (i) the ability to keep the undisclosed test dataset *safe and secure from malicious parties that could obtain an unfair advantage by somehow getting access to the test dataset and (ii) the ability to prove that the evaluation performed on such undisclosed test dataset is performed correctly.* Failure in the protection of this dataset and the evaluation phase could immediately and irreparably damage the reputation of FG-AI4GH.

In this short document, we propose a privacy-preserving approach for the above-mentioned benchmarking pipeline that ensures strong data confidentiality and accountability. Particularly, we describe how advanced privacy-enhancing technologies, such as homomorphic encryption and permissioned distributed ledgers, can be used as key enablers for reinforcing the feeling of trust that people will have in the whole process.

# 2. Homomorphic Encryption

Homomorphic encryption is a special form of public-key encryption that enables computations on encrypted values (see Figure 1). As with every public-key encryption scheme, it consists of two algorithms: an encryption algorithm and a decryption algorithm. Encryption and decryption can be seen as inverse operations. At one end, the encryption algorithm takes as input a plaintext message *m* and a public cryptographic key *K* and outputs a ciphertext *c*. On the other end, the decryption algorithm takes as input the ciphertext *c* and a secret cryptographic key *k* and outputs the original message *m*.

In general, the public key *K* is derived from a secret key *k* in such a way that the inverse operation is not possible. We can say that an encryption scheme is secure if the ciphertext does not reveal any information about the underlying message to anyone without the secret key *k* or, similarly, if anyone without the secret key *k* is unable to distinguish an encryption of a message *m0* from the encryption of another message *m1*.



Figure 1. High-level representation of homomorphic encryption enabling computations on encrypted data. Different keys are used to encrypt and decrypt messages.

In contrast to standard public-key encryption that requires decryption in order to carry out any computation on the original messages, homomorphic encryption allows to perform computations directly on the encrypted messages. In fact, homomorphically-encrypted messages can be combined together, thus generating an encrypted result that, when decrypted, yields the result of a meaningful operation (like addition or multiplication) on the original messages. For example, by using the homomorphic encryption scheme ElGamal on elliptic curves [2], also denoted as EC ElGamal, one can decrypt the sum of two ciphertexts and obtain the result of the sum of the two original messages.

Thanks to this property, homomorphic encryption schemes enable several use cases such as performing analytics in an untrusted environment (e.g., a public cloud) while ensuring the confidentiality of the data processed. Yet, it is important to note that homomorphic encryption introduces sometimes unpractical costs. For example, fully homomorphic encryption allows any computation to be carried out on the ciphertext but at the cost of unacceptable storage and computational overheads. Instead, partially homomorphic encryption has less flexibility than fully homomorphic encryption, as it allows only some specific computations to be carried out the ciphertext, but with reasonable storage and computation overheads. In this paper, we propose to use partially (but practical) homomorphic encryption for privacy-preserving benchmarking of AI-based models.

# 3. Proposed Privacy-Preserving Approach

Here, we outline our proposal for a privacy-preserving benchmarking pipeline that aims at achieve the following privacy and security goals in the malicious adversarial model:

* Absence of a single point of trust (or failure) in the benchmarking process
* End-to-end confidentiality protection of the undisclosed test dataset
* Correctness of the evaluation of submitted models
* Strong accountability and auditability of all the actions performed in the system

Given the international and collaborative nature of the effort of FG-AI4GH, it will be unlikely and undesired to have to trust only a single party or authority for the protection of the undisclosed test dataset. Such authority would represent a so-called “single point of failure” in the system as, if compromised, it could put the whole benchmarking process into jeopardy. For this reason, we propose to rely on the notion of “collective authority” that was initially introduced by Froelicher et al. [3] and has been recently applied to the medical field by Raisaro et al. [4] with the system MedCo (<https://medco.epfl.ch/>). A collective authority consists of a set of independent (and possibly heterogeneous) institutions (such as WHO, ITU, EPFL and Lausanne University Hospital) that decide to cooperate in order to collectively ensure the protection of the data and to be equal recipients of trust.

To achieve these notions of collective protection and trust decentralization, the collective authority makes use of distributed cryptographic techniques. During the initialization of the benchmarking platform, each institution in the collective authority needs to generate a pair of cryptographic keys (a public key and a secret key) and to collaborate with the other institutions in order to build a “collective” encryption key that is the result of the aggregation of all the single public keys. The collective encryption key can be used to encrypt the undisclosed test dataset by the means of a partially homomorphic cryptosystem. As such, the confidentiality of test dataset can be protected against malicious adversaries at storage, in transit and during computation as long as at least one of the institutions is not compromised. The encrypted dataset can be stored at any of the institutions forming the collective authority or even at an untrusted third party (e.g., a public cloud provider) with almost no risks. Indeed, an adversary, in order to decrypt the data, has to compromise all the institutions in the collective authority, steal the single secret keys and reconstruct the “collective” secret key that physically does not exist anywhere in its full form. In short: the larger is the collective authority, the more secure is the data. In addition to the generation of the collective encryption key, the collective authority will maintain a permissioned distributed ledger (also known as private blockchain) in order to have a distributed tamper-proof log that ensures accountability and auditability of all the actions performed during the benchmarking process.

In a nutshell, our privacy-preserving version of the benchmarking pipeline described in [1] can be summarized as follows:

1. After training a AI-based model on a combination of public and private data, a researcher will have to submit it to the evaluation platform together with his public key. This action will be immutably logged on the distributed ledger for auditability and accountability purposes. We note that the evaluation platform can be operated by any of the institutions in the collective authority.
2. The benchmarking platform will evaluate the submitted model on the encrypted undisclosed test dataset by leveraging the homomorphic properties of the cryptosystem and will generate an encrypted evaluation. The cleartext dataset will never be exposed as the evaluation of the model will happen in the encrypted domain. This step guarantees end-to-end confidentiality protection for the test dataset with no single point of trust (or failure).
3. For each operation carried out in the encrypted domain, a cryptographic zero-knowledge proof of correctness will be also generated and logged on the distributed ledger in order to be verified by the researcher. This step will guarantee strong accountability and transparency in the evaluation phase.
4. Before being published, the evaluation result undergoes a distributed re-encryption protocol across the collective authority in order to be re-encrypted from the collective public key to the researcher’s public key.
5. Finally, the result is immutably stored on the distributed ledger so that the researcher can visualize it after decryption. Optionally, the result can be directly published on the ledger in cleartext.

# 4. Conclusion and Next Steps

In this short document, we have outlined a potential privacy-preserving solution for securing the benchmarking pipeline put forward by WHO/ITU Focus Group on Artificial Intelligence for Health in its first white paper. Our goal is further explore this promising path and develop a first prototype in collaboration with the Lausanne University Hospitals, EPFL and its Center for Digital Trust (<https://c4dt.org/>) that will integrate collective homomorphic encryption from the UnLynx library [3] and distributed ledger technologies into a benchmarking platform such as <https://www.crowdai.org/>. The initial prototype will initially focus on linear models and will then evolve to cover a broader spectrum of AI-based algorithms.

# References

[1] Salathé M, Wiegand T, Wenzel M and Kishnamurthy R, *Focus Group on Artificial Intelligence for Health*, White paper <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FG-AI4H_Whitepaper.pdf>

[2] Bernstein DJ, Duif N, Lange T, Schwabe P, Yang B-Y. High-speed high-security signatures. J Cryptogr Eng. 2012 Sep 14;2(2):77–89.

[3] Froelicher D, Egger P, Sousa JS, Raisaro JL, Huang Z, Mouchet C, Ford B, Hubaux JP. UnLynx: a decentralized system for privacy-conscious data sharing. Proceedings on Privacy Enhancing Technologies. 2017 Oct 1;2017(4):232-50.

[4] Raisaro JL, Troncoso-Pastoriza J, Misbach M, Sousa JS, Pradervand S, Missiaglia E, Michielin O, Ford B, Hubaux JP. MEDCO: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. IEEE/ACM transactions on computational biology and bioinformatics. 2018 Jul 13.

[5] Camenisch J, Stadler M. Proof systems for general statements about discrete logarithms. Technical report/Dept. of Computer Science, ETH Zürich. 1997;260.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_