PhD proposal: *Exploring the Scope of Machine Learning using Homomorphic encryption under missing data in IoT/Cloud*

Supervisors: Prof. Samia Bouzefrane (ROC team) & Dr. Vincent Audigier (MSDMA team)

Introduction:

According to Gartner[1], 5.8 Billion Enterprise and Automotive IoT endpoints will be in use at the end of 2020 while Statistica[2] shows that IoT enablers solutions (such as Cloud, analytics, security) will reach 15 Billion of euros in the European Union market by 2025. However, these IoT devices have not enough resource capacity to process the data collected by their sensors making these devices vulnerable and prone to attack. To avoid processing data within the IoT devices, the trend is to outsource the sensed data to the Cloud that has both resourceful data storage and data processing. Nevertheless, the externalized data may be sensitive, and the users may lose privacy on the data content while allowing the cloud providers to access and possibly use these data to their own business. To avoid this situation and preserve data privacy in the Cloud datacenter, one possible solution is to use the fully homomorphic encryption (FHE) that assures both confidentiality and efficiency of the processing. In many smart environments such as smart cities, smart health, smart farming, industry 4.0, etc. where massive data are generated, there is a need to apply machine learning (ML) techniques, hence contributing to the decision making to act on the smart environment. Indeed, the challenging issue in this context is to adapt the ML approaches to apply them on encrypted data so that the decision taken on encrypted data can be reported on the cleartext data.

Context

A large amount of data is increasingly collected from different devices such as health devices, automotive equipment, smart home, industry 4.0. On the other hand, machine learning and statistical techniques are powerful tools for analyzing massive data [1]. However, ML on sensitive data (such as medical data) requires a special attention regarding privacy concerns and regulations. FHE enables computation over encrypted data [2, 3] preventing third parties (such as cloud providers) from accessing the cleartext data. This motivates the use of computational intelligence techniques such as ML on outsourced data.

Rivest et al. in [4] investigated the concept of computing over encrypted data called "privacy homomorphism", Gentry [5] introduced the first FHE and others [6-8] proposed more efficient schemes. While authors investigated privacy-preserving applications FHE [9, 10] for medical data either in Cloud or IoT context, others introduced ML models with FHE in medicine as surveyed in [1]. In 2017, authors in [11] trained logistic regression models on encrypted data by modifying the optimization function. Bost et al. [12] presented a privacy-preserving method for naive Bayes classification while several works consider FHE to achieve private classification of encrypted data over neural networks [13, 14, 15, 16, 17, 18].

Despite the existing works, as stated above, that deal with privacy preservation using ML models, to the best of our knowledge there is no research work that handles missing encrypted data. In fact, in many smart applications, we can have a broken device that cannot produce any data or any IoT device that has connection problem avoiding the reception by the Cloud provider of encrypted data from the data source.

[1] https://www.gartner.com/en/newsroom/press-releases/2019-08-29-gartner-says-5-8-billion-enterprise-and-automotive-io

[2] https://www.statista.com/statistics/686198/iot-solutions-market-in-the-european-union-eu/

Objective

This PhD thesis is a cooperative research work between two teams ROC and MSDMA of CEDRIC Lab. It aims at exploring the use of ML & FHE in smart applications where IoT devices collect sensitive data to outsource them on untrusted Cloud datacenter for computing thanks to ML models.

While neural networks are intensively studied with FHE, other models are less explored. The objective of this PhD thesis is twofold:

- First, we explore new other ML models with FHE that can be suitable in IoT/Cloud context to preserve privacy. The challenge is to carry out the appropriate modification of the ML model with FHE while considering a complete encrypted dataset in the Cloud server.
- Second, we assume missing encrypted data among the dataset the Cloud to design an appropriate solution that handles FHE as well as missing data.

Methodology

The methodology to adopt must follow the different steps:

- The research work must start with a state of the art to study the ML models that handle FHE.
- Based on the constraints of the smart applications, the PhD student must identify a supervised ML technique and propose an approach dealing with FHE. Based on benchmark datasets, this approach will be compared to the ML technique using cleartext data.
- The method will be then compared to the state-of-the-art supervised ML methods. Comparison will be in terms of security, scalability and accuracy.
- As a second phase, the PhD student will consider a scenario that is characterized by missing data to investigate suitable methods handling missing values in such settings. Matrix completion techniques [19, 20], multiple imputation methods [21, 22] or maximum likelihood approaches [23] will be investigated.
- Adapt the chosen method to handle FHE as well as missing data, test the proposed approach with several datasets.

Planning

1. During the 1st year:
   - 6 months: state of the art.
   - 6 months: first proposed approach ML & FHE.
   - First paper to submit in a conference based on this research work
2. During the 2nd year:
   - 5 months: Generalization and test of the model
   - 1 month: Second paper to submit to a journal
   - 6 months: Study of approaches with missing data.
3. During the 3rd year:
   - 6 months: Design of a solution with missing data and submit a journal paper
   - 6 months: write the PhD dissertation.

References

[1] Kahrobaei, Delaram, Wood, Alexander and Najarian, Kayvan (2020) Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics. ACM Computing Survey, April 2020. doi:10.1145/3394658.

[2] Craig Gentry. 2010. Computing Arbitrary Functions of Encrypted Data. Commun. ACM 53, 3 (March 2010), 97-105.

[3] V. Vaikuntanathan. 2011. Computing Blindfolded: New Developments in Fully Homomorphic Encryption. In 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science. 5-16.

[4] R. L. Rivest, A. Shamir, and L. Adleman. 1978. A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM 21, 2 (Feb. 1978), 120-126.

[5] C. Gentry, "Fully Homomorphic Encryption Using Ideal Lattices," in Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing, ser. STOC '09, 2009, pp. 169–178.

[6] Z. Brakerski and V. Vaikuntanathan, "Efficient Fully Homomorphic Encryption from (Standard) LWE." in FOCS, 2011, pp. 97–106.

[7] J. Fan and F. Vercauteren, "Somewhat Practical Fully Homomorphic Encryption." IACR Cryptology ePrint Archive, vol. 2012, p. 144, 2012.

[8] C. Gentry, A. Sahai, and B. Waters, "Homomorphic Encryption from Learning with Errors: Conceptually-Simpler, Asymptotically-Faster, Attribute-Based." in CRYPTO, ser. Lecture Notes in Computer Science, vol. 8042, 2013, pp. 75–92.

[9] S. Carpov, T. H. Nguyen, R. Sirdey, G. Constantino and F. Martinelli, "Practical Privacy-Preserving Medical Diagnosis Using Homomorphic Encryption," 2016 IEEE 9th International Conference on Cloud Computing (CLOUD), San Francisco, CA, 2016, pp. 593-599, doi: 10.1109/CLOUD.2016.0084.

[10] L. Jiang, L. Chen, T. Giannetsos, B. Luo, K. Liang and J. Han, "Toward Practical Privacy-Preserving Processing Over Encrypted Data in IoT: An Assistive Healthcare Use Case," in IEEE Internet of Things Journal, vol. 6, no. 6, pp. 10177-10190, Dec. 2019, doi: 10.1109/JIOT.2019.2936532.

[11] Andrey Kim, Yongsoo Song, Miran Kim, Keewoo Lee, and Jung Hee Cheon. 2018-10-11. Logistic regression model training based on the approximate homomorphic encryption. 11, 4 (2018-10-11), 83.

[12] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shai Goldwasser. 2015. Machine Learning Classification over Encrypted Data. Symposium on Network and Distributed System Security (NDSS).

[13] Ehsan Hesamifard, Hassan Takabi, Mehdi Ghasemi, and Rebecca N. Wright. 2018-01-01. Privacy-preserving Machine Learning as a Service. 2018, 3 (2018-01-01).

[14] Florian Bourse, Michele Minelli, Matthias Minihold, and Pascal Paillier. 2018. Fast Homomorphic Evaluation of Deep Discretized Neural Networks. In Advances in Cryptology - CRYPTO 2018 - 38th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2018, Proceedings, Part III. 483-512.

[15] Hervé Chabanne, Amaury de Wargny, Jonathan Milgram, Constance Morel, and Emmanuel Prouf. 2017. Privacy Preserving Classification on Deep Neural Network. Cryptology ePrint Archive, Report 2017⁄035. https:⁄⁄eprint.iacr. org⁄2017⁄035.

[16] RanGilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. In PMLR. 201-210

[17] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. 2018. GAZELLE: A Low Latency Framework for Secure Neural Network Inference. In 27th USENIX Security Symposium (USENIX Security 18). USENIX Association, Baltimore, MD, 1651-1669.

[18] Michele Minelli. Fully homomorphic encryption for machine learning. Cryptography and Security. PSL Research University, PhD thesis, 2018.

[19] Vincent Audigier, François Husson and Julie Josse. 2016. A principal components method to impute mixed data. Advances in Data Analysis and Classification, vol 10, no 1, pp 5-26.

[20] Trevor Hastie, Rahul Mazumder, Jason D. Lee and Reza Zadeh. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. 2015. J Mach Learn Res. Vol 16, pp 3367-3402.

[21] Donald B. Rubin, 1987. Multiple Imputation for Nonresponse in Surveys. Wiley.

[22] Vincent Audigier, Ian R. White, Shahab Jolani, Thomas Debray and Matthieu Resche-Rigon, 2018. Multiple imputation for multilevel data with continuous and binary variables. Statistical Science. Vol 33, number 2, pp 160-183.

[23] A. P. Dempster and N. M. Laird and D. B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. Vol 39, number 1, pp. 1-38.