AFR Individual PhD call 2023
Project Idea: Privacy-preserving and Interpretable Homomorphic Encryption-based Machine Learning

1. Introduction: Originality of the Research Project

In recent years, the development and application of machine learning (ML) has rapidly expanded, with significant implications for various fields. Consequently, concerns about data privacy and interpretability of ML models have become increasingly important. To address these concerns, researchers have developed various methods for privacy-preserving ML (PPML) [1-3] and Interpretable artificial intelligence (AI) [4-6], respectively. The aim of this research is to combine these two approaches into one coherent framework by developing an Interpretable AI method that can be applied to most promising PPML models, namely working on homomorphically encrypted data, thus enabling the development of AI systems that can be trusted by individuals and organizations alike.

Homomorphic encryption-based privacy-preserving machine learning (HE-PPML). Privacy-preserving machine learning has emerged as a critical research area, as organizations seek to protect sensitive data while still leveraging its value for machine learning applications. The recent draft regulation of the European Council [7] explicitly mentions the use of Privacy Enhancing Technologies as a preferred method versus traditional approaches which rely on direct data transfers.

Various PPML approaches exist, including fully homomorphic encryption (FHE), federated learning, and differential privacy, etc. FHE [8], or simply homomorphic encryption (HE), which is in the focus of our proposal, provides the most intuitive solution, namely, secure computation on encrypted data, allowing ML algorithms to be applied to sensitive data without decrypting it (Figure 1). HE contrasts with traditional encryption methods, which require data to be decrypted before any computation or prediction.

HE is a rapidly evolving field, with ongoing research aimed at improving its efficiency and practicality for use in real-world applications, including PPML [1, 3]. While the privacy of the data being processed is preserved, the output of the computation is also encrypted, and the only trusted party with the key may decrypt the final value.

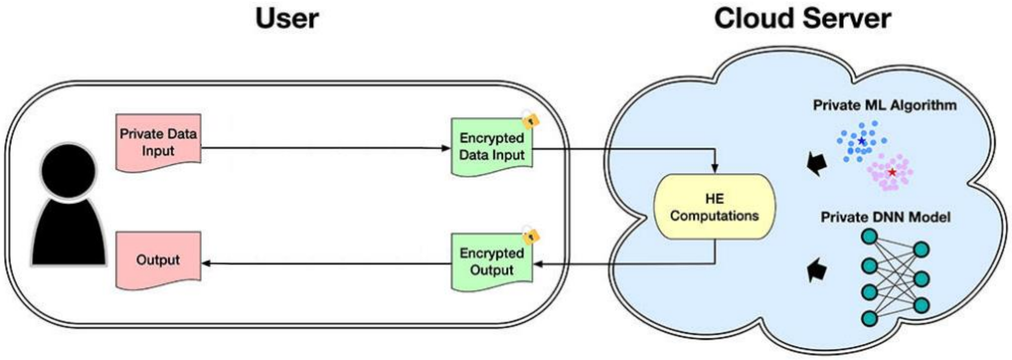

Figure 1: Workflow of HE-PPML

Interpretable AI. AI has made remarkable progress in recent years, but many models remain "black boxes" that are difficult to interpret and explain. This, in turn, reduces robustness of the models, i.e., the ability of a model to perform well on data that is different from the data on which it was trained. Interpretable AI has emerged as a critical research area to address this issue, providing methods for understanding how AI systems arrive at their decisions, which is required by the GDPR [9] by so-called "Right to explanation" [10].These methods can be divided into two groups [4]: inherently interpretable models (rule-based models, decision trees, linear models, etc.) and post-hoc interpretability, which is our focus here. Post-hoc interpretability aims at generating explanations for a specific model using an auxiliary model. Explicit examples are local explanations, which are relevant for complex models that cannot be interpreted globally. The only way is to explain each decision separately (locally) [5-6], such that, the predictions of the original models will be approximated in the local area of one data instance.

Originality of the Research Project. Privacy-preserving approaches have the drawback of increasing the complexity. They introduce an extra layer of data manipulation, which requires the employment of the new methodology of interpretation. As a result, applying Interpretable AI to PPML models has been the subject of extensive study over the last few years. The majority of such PPML models are not HE-based as for example, federated learning [2] or differential privacy [11]. Other studies simply looked at particular applications of Interpretable AI for particular circumstances utilizing HE-PPML, such as adapting the Grad-CAM attention map for face recognition [12]. They do not investigate the general framework and guidelines for using Interpretable AI on encrypted data. The absence of such investigations was primarily owe to the fact that the HE was previously computationally quite expensive. However, the remarkable progress in HE over the past 2-3 years makes it possible to apply FHE to more complex deep PPML models [1, 3] and, consequently, it is now feasible to conduct research on their interpretability.

2. Hypothesis, Research Objectives and Envisaged Methodology

The HE-PPML models raise an important challenge - the **lack of interpretability** of the resulting models. Because of encrypted data inputs, making interpretable explanations of how the model arrived at its decisions is even more challenging. Ensuring that the model makes accurate and unbiased decisions could become an issue. That **requires special method** to get **explainable decisions**.

To confirm our hypothesis, we suggest the following example. By now, local interpretable model-agnostic explanations (LIME) [5] is the pioneer method of explaining individual predictions made by black-box ML models. LIME focuses on developing local surrogate models that can explain a particular decision. The concept is straightforward and can be applied to different models working with different data types, including tabular, image, and text data. The workflow of LIME (Figure 2) is as follows:

1. Choosing the instance of interest for which is desired an explanation of the black-box model prediction.
2. Generating artificial data points by perturbations of the instance of interest.
3. Weighting perturbed points based on calculated distance from the chosen point.
4. Predicting the perturbed points with the black-box model.
5. Choosing an interpretable model (e.g., linear regression, decision tree, etc.).
6. For the local synthetic dataset, training a weighted, interpretable model and use it for explanation. In the case of linear regression, explaining is outputting the weights of the linear regression which indicate feature importance.

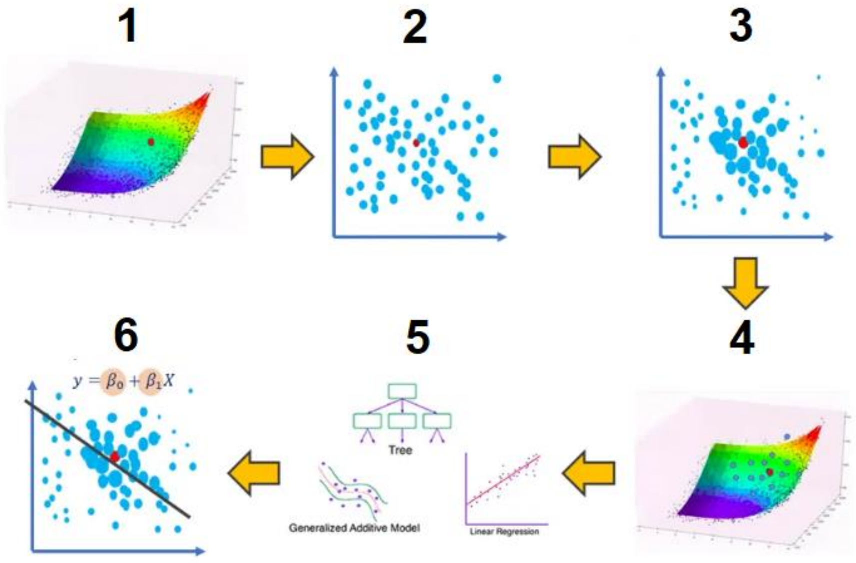

Figure 2: Workflow of LIME method

Now, assume the HE-PPML scenario in which the input data instance (for which prediction explanation is desired) is encrypted. In this case, perturbations will be in unencrypted format and data input - in encrypted format. For calculating the distance between perturbations (blue points on Figure 2) and the initial instance (red point), it requires them to be in the same format, which is not feasible. The reason is that in modern HE schemes, like FHE over the Torus (TFHE) [13], which can be used for PPML, the secret key for both encryption and decryption is required. Nevertheless, steps 2, 3, and 6 are impractical because the model is kept on the cloud server and the secret key is kept on the user side.

Our goal is to **assess the interpretability of ML models** when the **data is encrypted**. Furthermore, we are going to **identify optimal approaches** for creating interpretable ML models that operate on homomorphically encrypted data and to **examine the trade-offs between privacy, interpretability, and model accuracy** in PPML using homomorphic encryption. We **investigate the practical applications of homomorphic encryption** in real-world scenarios for **ensuring data privacy** while allowing interpretable machine learning.

The workflow of the PhD project can be decomposed into several stages:

1. Selection of a shortlist of interpretable AI methods from the literature and evaluation of their performance for HE-PPML using benchmarks as it was done in [14].
2. Developing a framework to achieve interpretability in HE-PPML.

   One approach is to use additional cryptographic primitives to ensure desired properties, in a similar way to [15], where the authors achieved IND-CCA2 security property by incorporating Zero-Knowledge Proofs (ZKP) into the Public-key encryption scheme. In our research ZKP can give masked generalized information for information model in order create explanation, for example where approximately the point is located or with noise.

   Another approach is to alter the protocols, which often require the model holder to produce the full explanation. We can offer users to create these explanations for themself. For instance, for LIME, it can work in such a way: the users generate synthetic data instances by themselves and builds linear regression on their own. The cloud server side only provides predictions.
3. The final stage of the experimental part project will involve experiments with the developed framework and simulated real-world problems, such as finance, automotive industry, etc. We will compare obtained results with existing methods for unencrypted data. This will demonstrate the practical value and impact of the research, as well as identify any potential challenges or limitations of the framework.

3. Expected Outcomes / Impact

Fundamental impact. The expected fundamental outcome of our research project is to **enhance the robustness of HE-PPML models**. We will achieve that by creating a special Interpretable AI framework, which can be used by models working on homomorphically encrypted data. By understanding the reasoning behind the output of the AI system, researchers and practitioners can identify and rectify any biases in the model. Using interpretable AI methods will ultimately lead to more accurate and reliable results, making HE-PPML more resilient.

Industrial application impact. The expected industrial application impact is to enable the development of HE-PPML models that can **meet the requirements of the GDPR** by providing transparency and accountability in the decision-making process. The existence of transparent explanations for the predictions enables users to understand how the model arrived at its conclusion. Consequently, it will push the development of **HE-PPML** that **can be used in real-world settings**, such as clinical decision support, automotive and credit risk assessment systems, etc. For example, it will change the way of communication of clients with Luxembourgish banks in significantly more private way.

4. Explanations on the management of ethical issues and data protection

There are no specific ethical issues.

There are also no data-protection issues. Our research project will involve the use only of artificially generated datasets to develop and test our ML models. This ensures that we will manage data protection issues appropriately and in compliance with GDPR guidelines and established UL internal processes.

Nevertheless, a Data Management Plan (DMP) will be created when starting the project as requested by the FNR rules and guidelines.

Bibliography

[1] M. Kumar, W. Zhang, L. Fischer и B. Freudenthaler, «Membership-Mappings for Practical Secure Distributed Deep Learning,» *IEEE Transactions on Fuzzy Systems,* 2023.

[2] T. Budig, S. Herrmann и A. Dietz, «Trade-offs between privacy-preserving and explainable machine learning in healthcare,» в *Seminar Paper, Inst. Appl. Informat. Formal Description Methods (AIFB), KIT Dept. Econom. Manage., Karlsruhe, Germany*, 2020.

[3] I. Chillotti, M. Joye и P. Paillier, «Programmable bootstrapping enables efficient homomorphic inference of deep neural networks,» в *Cyber Security Cryptography and Machine Learning: 5th International Symposium, CSCML 2021, Be'er Sheva, Israel, July 8–9, 2021, Proceedings 5*, 2021.

[4] R. Moraffah, M. Karami, R. Guo, A. Raglin и H. Liu, «Causal interpretability for machine learning-problems, methods and evaluation,» *ACM SIGKDD Explorations Newsletter,* т. 22, p. 18–33, 2020.

[5] M. T. Ribeiro, S. Singh и C. Guestrin, «" Why should i trust you?" Explaining the predictions of any classifier,» в *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[6] S. M. Lundberg и S.-I. Lee, «A unified approach to interpreting model predictions,» *Advances in neural information processing systems,* т. 30, 2017.

[7] Proposal for a regulation of the European Parliament, of the Council on European statistics on population, amending Regulation (EC) No 862/2007 housing, repealing Regulations (EC) No 763/2008, and (EU) No 1260/2013. European Commission.

[8] C. Gentry, «Fully homomorphic encryption using ideal lattices,» в *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009.

[9] Regulation (EU) 2016/679 of the European Parliament, of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data, on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1. European Commission.

[10] S. Wachter, B. Mittelstadt и C. Russell, «Counterfactual explanations without opening the black box: Automated decisions and the GDPR,» *Harv. JL & Tech.,* т. 31, p. 841, 2017.

[11] F. Harder, M. Bauer и M. Park, «Interpretable and differentially private predictions,» в *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[12] D. Franco, L. Oneto, N. Navarin и D. Anguita, «Toward learning trustworthily from data combining privacy, fairness, and explainability: an application to face recognition,» *Entropy,* т. 23, p. 1047, 2021.

[13] I. Chillotti, N. Gama, M. Georgieva и M. Izabachène, «TFHE: fast fully homomorphic encryption over the torus,» *Journal of Cryptology,* т. 33, p. 34–91, 2020.

[14] S. Hooker, D. Erhan, P.-J. Kindermans и B. Kim, «A benchmark for interpretability methods in deep neural networks,» *Advances in neural information processing systems,* т. 32, 2019.

[15] M. Naor и M. Yung, «Public-key cryptosystems provably secure against chosen ciphertext attacks,» в *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, 1990.