Literature Review of Secure Privacy-Preserving Inference

Artem Grigor[1,2]

[1] University Colledge London, UK
artem.grigor.23@ucl.ac.uk
[2] Confidenti, UK
artem@confidenti.ai

Abstract. In the digital age, ensuring the privacy and security of machine learning (ML) systems is not merely a necessity but a fundamental prerequisite for the field's growth and societal acceptance. One particular area of today's interest is development of secure ML Inference schemes suitable for adversarial settings, which can guarantee privacy of analysed data, secrecy of the used ML model and integrity of the results. This area, originally largely overlooked in favour of Privacy-Preserving Training, today finally is gaining the well deserved attention.

The research of Secure Privacy-Preserving Inference has oscillated between two distinct paradigms: cryptography-focused and machine learning security-focused approaches. The former adapts cryptographic schemes to machine learning (ML) computations, often overlooking ML's unique security and privacy requirements. The latter emphasizes empirical security enhancements while largely avoiding information security and cryptography principles, leading to repeated vulnerabilities discovered in ML models. In this survey we explore each both of these approaches in detail, understand how they contribute to Secure Privacy-Preserving Inference and attempt to bridge insights from both domains, suggesting a route to a comprehensive approach of achieving practical, real-world Secure Privacy-Preserving Inference lies on the intersection of the two approaches.

1 Introduction

In our current digital era, we have the unparalleled capability to generate, collect, and analyze vast quantities of data, unlocking significant benefits across multiple domains [27]. However, zettabytes of private data with the potential to transform lives remain largely untapped due to prevailing concerns over providing adequate privacy and security guarantees [12, 26][3].

Recent advancements in Differential Privacy [16], Federated Learning [30], and Synthetic Data generation [40] have facilitated practical Privacy-Preserving Machine Learning (ML) model training on distributed private data without explicit data sharing [7] with some security problems left to solve [25].

However, after we obtain the trained models using the techniques above, we now face a challenge of deploying the models to work securely with real private data, in a process called ML inference. And despite the advancements above, there exists a critical gap of knowledge , with no satisfactory solution present, as we argue bellow.

This, accompanied by a growing interest in running ML inference [4], highlights a pressing void in the current landscape: lack of schemes **to perform efficient, privacy-preserving and tamper-resistant inference on private data within adversarial environments** . Throughout this survey, we will refer to such schemes as **Secure Privacy-Preserving Inference** and we will explore different efforts done on the path to achieve Secure Privacy-Preserving Inference, which we will attempt to classify in groups and highlight correct security limitations of them.

1.1 Layout

The structure of the following sections is as follows:

1. First, we define the security settings of the Secure Privacy-Preserving Inference and define some terminology used throughout the paper.
2. Next, we explore background work done on the topic and suggest some promising areas of future research. We split the literature review into two major sections: Cryptography Focused Solutions 2 and Machine Learning Security-focused Solutions 3.
3. Finally, we summarise the current state of Secure Privacy-Preserving Inference and suggest areas we believe future research should focus to make Secure Privacy-Preserving Inference possible.

[3] Data Privacy needs to be combined with Data Utility - `https://www.prospectmagazine.co.uk/essays/52612/who-needs-digital-privacy`

[4] AI: Nvidia Focused on Inference - `https://medium.com/@mparekh/ai-nvidia-focused-on-inference-0ebf167238a0`

.

1.2 Security Model

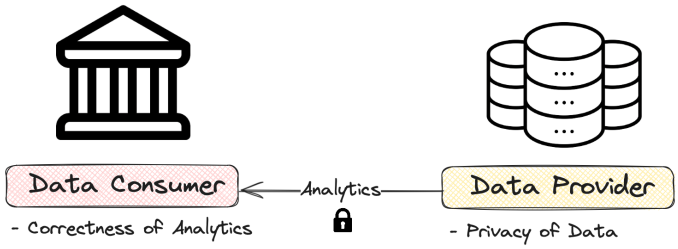

Fig. 1: Setting: Two mutually distrusting entities interacting with each other.

The security model for Secure Privacy-Preserving Inference outlines the interactions and security requirements between two mutually distrusting entities:

1. **Data Provider** - An entity possessing sensitive, potentially third-party data, aims to protect the data's confidentiality while utilizing it in a privacy-preserving manner to extract benefits from the data.

   Example: Individuals could serve as Data Providers, with their smartphone texts and photos acting as sensitive data. Utilizing such data could, for instance, improve the accuracy of credit scores or reduce insurance premiums, showcasing the benefits of secure data usage [18, 9].

2. **Data Consumer** - An entity that analyzes the Data Provider's information without necessarily compromising its privacy. It is vital for the Data Consumer to ensure the analytics results' accuracy and integrity are safeguarded from manipulations by the Data Provider.

   Example: Banks and Insurance Companies, acting as Data Consumers, could analyze clients' personal data to more accurately predict loan default risks [23] or health risks. This facilitates a more precise risk evaluation, illustrating the real-world benefits of accessing personal data.This facilitates a more precise risk evaluation, illustrating the real-world benefits of accessing personal data.

2 Cryptography-focused Solutions

Cryptography-focused solutions for Secure Privacy-Preserving Inference vary in defining the party that will perform the majority of computations over the personal data,

which we refer to as computational responsibility. We categorize the existing solutions based on the notion of computational responsibility and examine the state of these solutions, highlighting potential security and efficiency issues.

2.1 Bringing Data to ML

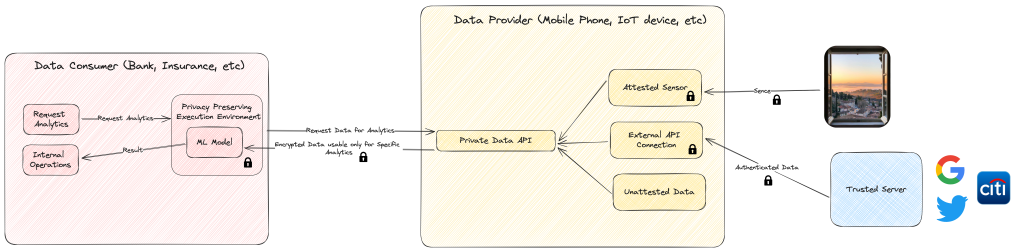

Fig. 2: Bringing Data to the ML Model.

This method is characterised by Data Consumer bearing most of the computation responsibility. It involves transferring data securely from the Data Provider to the Data Consumer, who then performs Secure Privacy Preserving Inference over the data. Techniques like Fully Homomorphic Encryption (FHE) [31, 4] and Trusted Execution Environments (TEE) [35], as well as the mix between them [37] aim to protect data during processing. However, challenges with FHE's practicality [33] and TEE's privacy guarantees [24, 39] suggest limitations, particularly in ensuring data deletion after it has been used. The inability to provably delete data without continuous auditing [39] raises concerns about the Data Consumer potentially retaining and later accessing the encrypted data [5].

There is also no safeguard against the Data Provider submitting incorrect or adversarial data as well as guarantee that the Data Consumer is not running a malicious models over the data. However, all the Cryptography-focused solutions as of today avoid these issues, and they have only been explored in the ML Security focused solutions, which we will cover this in the 3 section. Yet, we believe it should be more or less straightforward to port the solutions for such problems to integrate with the cryptography focused security solutions.

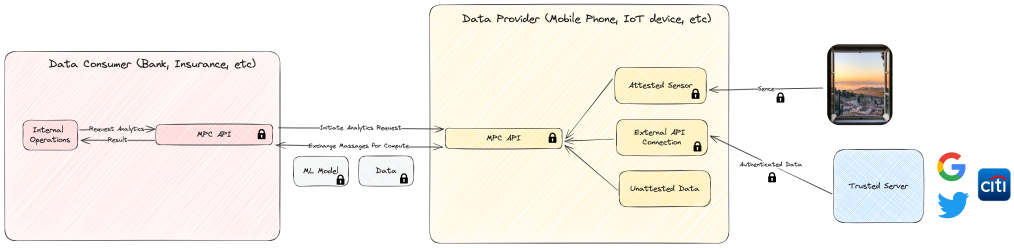

Fig. 3: Ongoing Interaction between Data Provider and Data Consumer.

2.2 Continuous Interactions

Featuring ongoing communication exchanges, this approach utilizes Secure Multi-Party Computation (SMPC) protocols [21] to balance computational loads. Here computation responsibility lies on both Data Consumer and Data Provider. While SMPC reduces some inefficiencies [4], it introduces performance penalties and continuous communication challenges, which might be unacceptable for mobile devices with bad coverage. Additionally, there is still a potential for Data Consumers to store interaction data for future decryption.

2.3 Bringing ML to Data

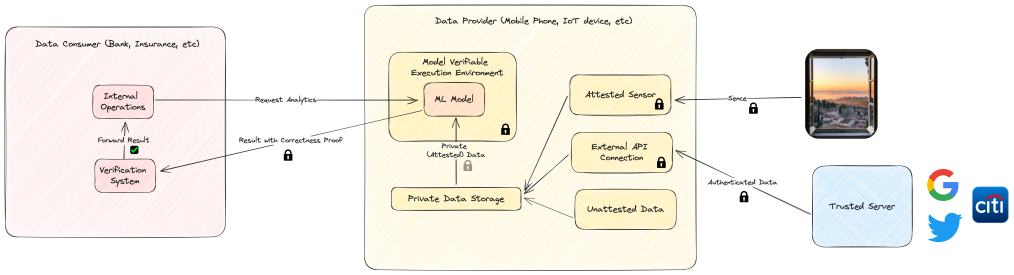

Fig. 4: Bringing the ML Model to Data.

The "Bringing ML to Data" strategy involves the Data Consumer sending the ML model to the Data Provider, who then runs the model locally on their dataset and returns the results. In this setting the Data Provider bears all the computational responsibility. This method inherently protects the Data Provider's sensitive

[5] NSA stores encrypted data until it can be cracked - https://www.techdirt.com/2013/06/21/nsa-has-convinced-fisa-court-that-if-your-data-is-encrypted-you-might-be-terrorist-so-itll-hang-onto-your-data

information by limiting the amount of data that is transmitted and avoiding to send raw data anywhere, thereby maintaining privacy by default. However, this approach requires the Data Consumer to trust the Data Provider's integrity in accurately executing the model and providing genuine results.

To address potential concerns of dishonesty and to assure the Data Consumer of the computation's integrity, verifiable computation schemes [41] are utilized. These schemes allow the Data Provider to offer cryptographic proof that the model was executed correctly and that the results are authentic. This development, known as Verifiable Machine Learning [38], is gaining importance as ML models are increasingly applied in sensitive sectors like healthcare, finance, and legal decision-making [19, 36].

The primary method for verifying ML model outputs involves the use of zk-SNARKs (Zero-Knowledge Succinct Non-Interactive Argument of Knowledge) to construct computational circuits that mimic the architecture of the ML model [14]. These circuits provide a cryptographic means to attest to the accuracy of inferences, catering to the verification of the model's evaluation. Despite active research to scale verifiable computations to match the complexity and size of contemporary ML models [38], significant challenges remain. Existing methods struggle to efficiently handle large-scale models [34], which today exceed gigabytes in size, leading to substantial computational costs for Data Providers, those with constrained resources, such as IoT devices. However, we want to point out there is an active interest in the area, such as a recent proposal to drastically scale verification of ML models using a clever Hint's method to verify larger models using smaller models [1].

Thus we believe that the approach of Bringing ML to Data, reinforced by verifiable computations, appears to us to be the most promising among existing methods. As the field of verifiable computations and non-interactive proofs attracts significant research and investment, highlighting the potential for future efficiency breakthroughs [38] [6]. Additionally, with edge devices becoming increasingly powerful [7], the capacity for conducting extensive computations locally is rapidly improving. This progress makes the "Bringing ML to Data" model not just feasible, but ideally suited for deployment in real-world scenarios.

[6] VC firm opens a cryptography department - https://www.coindesk.com/tech/2023/08/10/vc-firm-a16z-wades-into-crypto-tech-research-with-zk-projects-jolt-and-lasso/

[7] Your Phone is more powerful than your PC - https://insights.samsung.com/2021/08/19/your-phone-is-now-more-powerful-than-your-pc-3/

3 Machine Learning Security-focused Solutions

While cryptography-focused solutions provide robust frameworks for Secure Privacy-Preserving Inference, they often treat Machine Learning (ML) models as black boxes, neglecting the considerable risks posed by model exploitation, such as backdoor-ed models [13], adversarial inputs [11] and model jail breaking [42], to name a few. This oversight can lead to vulnerabilities that compromise the integrity of Secure Privacy-Preserving Inference systems from within, bypassing the security provided by cryptography focused solutions. The following discussion highlights key areas of concern, and explores some mitigation strategies. We find that the state of ML Security today is far from being comprehensive and emphasize the importance of a more nuanced approach to ML Security.

3.1 Malicious Data Provider

Adversarial ML Malicious Data Providers pose a significant threat by manipulating model outputs through the injection of specially crafted adversarial data. This technique can cause models to produce adversary controlled results, bypassing the protective measures of cryptography-based solutions like verifiable computations, which do not account for internal logic issues of ML models [11]. Addressing this challenge requires a robust ML security framework capable of detecting and neutralizing adversarial data threats. One common countermeasure to make crafting adversarial data harder is to keep the model close sourced, however it has been shown that it is possible to extract such models given enough queries [28] or be attacked directly [29]. Another potential countermeasure is the use of attested clean data, certified by third parties or sensor attestations, though this area remains underexplored in current research, and we believe it can still be susceptible for some attacks [17, 42] . To make matters worth, we believe that there might be no satisfactory solution as long as we are using Deep Neural Networks, and we might need to resort to alternative solutions as an only way to make it secure.

ML Extraction Additionally, malicious Data Providers may seek to access or infer proprietary information embedded within the Data Consumer's model, aiming to reverse-engineer the model for competitive advantages [8] or to access private training data [28]. Solutions like Trusted Execution Environments (TEE) and Fully Homomorphic Encryption (FHE) have been proposed to shield the models' parameter, but

[8] Apple NeuroHash Model reverse-engineered - `https://www.schneier.com/blog/archives/2021/08/apples-neuralhash-algorithm-has-been-reverse-engineered.html`

vulnerabilities in machine learning still allow to perform parameter extraction [3]. This situation underscores the imperative for more sophisticated security measures that extend beyond TEE and FHE, advocating for ongoing research in this domain [2].

3.2 Malicious Data Consumer

Malicious and Backdoored Models On the flip side, Malicious Data Consumers aim to breach data privacy by exploiting models to extract Data Provider information. Direct methods include deploying models that simply output received data rather than performing intended computations. Mitigating such threats may require rigorous audits and transparency regarding model weights and architectures. Emerging research has revealed techniques for embedding covert triggers in model parameters or architectures, triggering unauthorized functions like data exfiltration upon detecting specific inputs [15, 6], refereed to as backdoor. These revelations highlight the urgent need for enhanced security protocols to prevent such clandestine operations. We note that there is a close correlation between backdoored models and adversarial examples, as the only difference of a backdoored model is that an adversarial example is embedded during training[32], rather than found during model examination. Thus there still exists an open question if any defense is possible with current model architectures.

Query Abuse Finally, excessive querying by Data Consumers can jeopardize data privacy by performing reconstruction attack [20]. Local Differential Privacy (DP)[22] has been touted as a solution for safeguarding individual data points within datasets , [10], but its applicability to inference processes is still unclear. Furthermore, the current conservatives of DP estimation techniques leads it to be too impractical to be used [8]. There is a critical need for research aimed at optimizing DP measures, possibly through dynamic adjustment of the Privacy Budget based on previous queries, to strike a better balance between privacy preservation and utility [5].

4 Conclusion

Addressing the security challenges posed by both malicious Data Providers and Consumers is crucial for the development of Secure Privacy-Preserving Inference systems. While cryptography-based solutions lay a strong foundation, the unique threats within the ML landscape demand specialized security measures. A comprehensive approach, incorporating robust ML security frameworks and ongoing research

into advanced cryptographic and verification techniques, is essential for safeguarding against the sophisticated tactics employed by adversaries in the digital age.

Bibliography

[1] Artem Grigor, Anton Kravchenko, G.W.: Efficient verification framework for large-scale machine learning models. https://github.com/ElusAegis/papers/blob/main/ML-Verifier-Hints.pdf (April 2024), available online at https://github.com/ElusAegis/papers/blob/main/ML-Verifier-Hints.pdf

[2] Atrey, A., Sinha, R., Mitra, S., Shenoy, P.J.: SODA: protecting proprietary information in on-device machine learning models. CoRR **abs/2312.15036** (2023). https://doi.org/10.48550/ARXIV.2312.15036, https://doi.org/10.48550/arXiv.2312.15036

[3] Atrey, A., Sinha, R., Sarkhel, S., Mitra, S., Arbour, D., Maharaj, A., Shenoy, P.: Towards preserving server-side privacy of on-device models. In: Companion Proceedings of the Web Conference 2022. p. 282–285. WWW '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3487553.3524257, https://doi.org/10.1145/3487553.3524257

[4] Azraoui, M., Bahram, M., Bozdemir, B., Canard, S., Ciceri, E., Ermis, O., Masalha, R., Mosconi, M., Önen, M., Paindavoine, M., Rozenberg, B., Vialla, B., Vicini, S.: SoK: Cryptography for Neural Networks, pp. 63–81. Springer International Publishing, Cham (2020), https://doi.org/10.1007/978-3-030-42504-3_5

[5] Bai, Y., Yang, G., Xiang, Y., Wang, X.: Generalized and multiple-queries-oriented privacy budget strategies in differential privacy via convergent series. Security and Communication Networks **2021**, 1–17 (12 2021). https://doi.org/10.1155/2021/5564176

[6] Bober-Irizar, M., Shumailov, I., Zhao, Y., Mullins, R., Papernot, N.: Architectural backdoors in neural networks (2022)

[7] Chen, Q., Xiang, C., Xue, M., Li, B., Borisov, N., Kaarfar, D., Zhu, H.: Differentially private data generative models (2018)

[8] Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Huang, Y., Jagielski, M., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., Papernot, N., Rogers, R., Shen, M., Song, S., Su, W., Terzis, A., Thakurta, A., Vassilvitskii, S., Wang, Y.X., Xiong, L., Yekhanin, S., Yu, D., Zhang, H., Zhang, W.: Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. Harvard Data Science Review **6**(1) (jan 16 2024), https://hdsr.mitpress.mit.edu/pub/sl9we8gh

[9] Djeundje, V.B., Crook, J., Calabrese, R., Hamid, M.: Enhancing credit scoring with alternative data. Expert Systems with Applications **163**, 113766 (2021). https://doi.org/https://doi.org/10.1016/j.eswa.2020.113766, https://www.sciencedirect.com/science/article/pii/S095741742030590X

[10] Ebadi, H., Sands, D., Schneider, G.: Differential privacy: Now it's getting personal. In: Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages. p. 69–81. POPL '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2676726.2677005, https://doi.org/10.1145/2676726.2677005

[11] Fenaux, L., Kerschbaum, F.: Sok: Analyzing adversarial examples: A framework to study adversary knowledge (2024)

[12] García-Gasco Romero, M.: Personal Data: The New Black Gold, pp. 171–182. Springer International Publishing, Cham (2021), https://doi.org/10.1007/978-3-030-67973-6_12

[13] Goldwasser, S., Kim, M.P., Vaikuntanathan, V., Zamir, O.: Planting undetectable backdoors in machine learning models (2022)

[14] Groth, J.: On the size of pairing-based non-interactive arguments. In: Fischlin, M., Coron, J.S. (eds.) Advances in Cryptology – EUROCRYPT 2016. pp. 305–326. Springer Berlin Heidelberg, Berlin, Heidelberg (2016)

[15] Gu, T., Dolan-Gavitt, B., Garg, S.: Badnets: Identifying vulnerabilities in the machine learning model supply chain (2019)

[16] Ji, Z., Lipton, Z.C., Elkan, C.: Differential privacy and machine learning: a survey and review (2014)

[17] Khachaturov, D., Gao, Y., Shumailov, I., Mullins, R., Anderson, R., Fawaz, K.: Human-producible adversarial examples (2023)

[18] Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. IEEE Communications Magazine **48**(9), 140–150 (2010). https://doi.org/10.1109/MCOM.2010.5560598

[19] Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.: Trustworthy ai: From principles to practices (2022)

[20] Liu, S., Wang, Z., Lei, Q.: Data reconstruction attacks and defenses: A systematic evaluation (2024)

[21] Long, Y., Gangwani, T., Mughees, H., Gunter, C.: Distributed and secure ml with self-tallying multi-party aggregation (2018)

[22] Mahawaga Arachchige, P.C., Bertok, P., Khalil, I., Liu, D., Camtepe, S., Atiquzzaman, M.: Local differential privacy for deep learning. IEEE Internet of Things Journal **7**(7), 5827–5842 (2020). https://doi.org/10.1109/JIOT.2019.2952146

[23] Meier, S., Sprenger, C.: Impatience and credit behavior: Evidence from a field experiment. Federal Reserve Bank of Boston, Working Papers (01 2007). https://doi.org/10.2139/ssrn.982398

[24] Muñoz, A., Ríos, R., Román, R., López, J.: A survey on the (in)security of trusted execution environments. Computers and Security **129**, 103180 (2023). https://doi.org/https://doi.org/10.1016/j.cose.2023.103180, https://www.sciencedirect.com/science/article/pii/S0167404823000901

[25] Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. ArXiv **abs/1812.00910** (2018), https://api.semanticscholar.org/CorpusID:54444175

[26] Niu, C., Wu, F., Tang, S., Ma, S., Chen, G.: Toward verifiable and privacy preserving machine learning prediction. IEEE Transactions on Dependable and Secure Computing **19**(3), 1703–1721 (2022). https://doi.org/10.1109/TDSC.2020.3035591

[27] Nuccio, M., Guerzoni, M.: Big data: Hell or heaven? digital platforms and market power in the data-driven economy. Competition & Change **23**(3), 312–328 (2019). https://doi.org/10.1177/1024529418816525, https://doi.org/10.1177/1024529418816525

[28] Oh, S.J., Augustin, M., Schiele, B., Fritz, M.: Towards reverse-engineering black-box neural networks (2018)

[29] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning (2017)

[30] Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Úlfar Erlingsson: Scalable private learning with pate (2018)

[31] Podschwadt, R., Takabi, D., Hu, P., Rafiei, M.H., Cai, Z.: A survey of deep learning architectures for privacy-preserving machine learning with fully homomorphic encryption. IEEE Access **10**, 117477–117500 (2022). https://doi.org/10.1109/ACCESS.2022.3219049

[32] Shumailov, I., Shumaylov, Z., Kazhdan, D., Zhao, Y., Papernot, N., Erdogdu, M.A., Anderson, R.: Manipulating sgd with data ordering attacks (2021)

[33] Stoian, A., Frery, J., Bredehoft, R., Montero, L., Kherfallah, C., Chevallier-Mames, B.: Deep neural networks for encrypted inference with tfhe. Cryptology ePrint Archive, Paper 2023/257 (2023), https://eprint.iacr.org/2023/257, https://eprint.iacr.org/2023/257

[34] Team, M.L.: The cost of inteligence: Proving machine learning inference with zero knoweldge. Online (2023), https://github.com/Modulus-Labs/Papers/blob/master/Cost_Of_Intelligence.pdf, accessed: 2024-04-02

[35] Truong, J.B., Gallagher, W., Guo, T., Walls, R.J.: Memory-efficient deep learning inference in trusted execution environments (2021)
[36] Vice: A judge just used chatgpt to make a court decision. https://www.vice.com/en/article/k7bdmv/judge-used-chatgpt-to-make-court-decision (2023), accessed: 2023-09-29
[37] Wang, Q., Zhou, L., Bai, J., Koh, Y.S., Cui, S., Russello, G.: Ht2ml: An efficient hybrid framework for privacy-preserving machine learning using he and tee. Computers and Security **135**, 103509 (2023). https://doi.org/https://doi.org/10.1016/j.cose.2023.103509, https://www.sciencedirect.com/science/article/pii/S0167404823004194
[38] Xing, Z., Zhang, Z., Liu, J., Zhang, Z., Li, M., Zhu, L., Russello, G.: Zero-knowledge proof meets machine learning in verifiability: A survey (2023)
[39] Yang, C., Liu, Y., Zhao, F., Zhang, S.: Provable data deletion from efficient data integrity auditing and insertion in cloud storage. Computer Standards & Interfaces **82**, 103629 (2022). https://doi.org/https://doi.org/10.1016/j.csi.2022.103629, https://www.sciencedirect.com/science/article/pii/S0920548922000101
[40] Yoon, J., Jordon, J., van der Schaar, M.: PATE-GAN: Generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=S1zk9iRqF7
[41] Yu, X., Yan, Z., Vasilakos, A.V.: A survey of verifiable computation. Mobile Networks and Applications **22**(3), 438–453 (06 2017), https://www.proquest.com/scholarly-journals/survey-verifiable-computation/docview/1908019401/se-2, copyright - Mobile Networks and Applications is a copyright of Springer, 2017; Last updated - 2023-11-30
[42] Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models (2023)