A MACHINE LEARNING APPROACH TO OPTIMIZE EFFICIENCY AND PRIVACY IN SECURE MULTI-PARTY COMPUTATION

Anuradha Rajendra More
Research Scholar
Shri JJT University,
Rajasthan.

Dr. Dand Hiren Jayanti Lal
Professor
Guide
Shri JJT University,
Rajasthan.

Dr. Santosh T. Jagtap.
Professor
Co-Guide
Shri JJT University,
Rajasthan.

ABSTRACT

Secure Multi-Party Computation (MPC) protocols allow a set of mutually-distrusting parties to jointly evaluate a commonly known function over their inputs, while maintaining correctness of the outputs and the security of their inputs. Privacy-preserving machine learning (PPML) and Secure Multi-party Computation (MPC) has gained momentum in the recent past. As its deployment increasingly depends on data from multiple entities, ensuring privacy for these contributors becomes paramount for the integrity and fairness of machine learning endeavors. We substantiate our theoretical claims through improvement in benchmarks of the aforementioned algorithms when compared with the current best framework ABY3. All the protocols are implemented over a 64-bit ring in LAN and WAN. Albeit its potential, the practicality of MPC is hindered by the difficulty to implement applications on top of the underlying cryptographic protocols. This is because their manual construction requires expertise in cryptography and hardware design. The latter is required as functionalities in MPC are commonly expressed by Boolean and Arithmetic circuits, whose creation is a complex, error-prone, and time-consuming task. We begin with an introduction into compilation and optimization of circuits with minimal size, which is required for constant round MPC protocols over Boolean circuits, such as Yao's Garbled Circuits protocol.

Keywords: *Secure Multi-Party Computation (MPC) protocols, Privacy-preserving machine learning (PPML), cryptographic protocols, LAN and WAN, mutually-distrusting.*

INTRODUCTION

Private information is susceptible to risks linked to data breaches, financial damage, reputational harm, and legal repercussions, making security a stakeholder's top priority. Privacy is another essential factor. Data frequently contains sensitive or confidential data, and parties need to guarantee that the privacy privileges of users are honoured. Compliance with confidentiality laws like the General Data Protection Regulation (GDPR) makes data sharing even more difficult, as organisations have to negotiate complex legal structures to protect privacy. Simple pattern recognition and learning from single data sources have given means to more complicated processes such as collaborative machine learning across domains and Generative AI which relies on foundational models, pre-trained models trained on vast amounts of data. These developments have revolutionised the generation of enterprise databased insights. Machine Learning insights can be made precise, trustworthy, and efficient by ensuring that the data is coming from various stakeholders to cover variety, volume, and distribution. Many a time, the machine learning models fail when applied to unforeseen data as a result of a lack of diverse perceptions and diverse data sources to train with. It is essential to recognise that the data employed to generate these insights is frequently restricted to the business's data. Therefore,

whenever these models are utilised in various organisational contexts, their precision, effectiveness, and consistency may suffer. Organisations can surmount these obstacles and improve the efficacy of machine learning findings by adopting a free data exchange between multiple stakeholders, to contribute to a diverse dataset. The extended dataset permits machine learning models to learn from a broader range of examples and scenarios, resulting in more precise and trustworthy insights. Sharing data also enables the detection of trends and patterns that might not be evident when data are analysed in isolation. Combining data from multiple constituents, organisations may discover concealed insights and obtain a greater awareness of complex problems. Data sharing fosters accountability and transparency in machine learning processes.

LITERATURE REVIEW

Massimo Piccardi (2024) Machine learning is a powerful technology for extracting information from data of diverse nature and origin. As its deployment increasingly depends on data from multiple entities, ensuring privacy for these contributors becomes paramount for the integrity and fairness of machine learning endeavors. This review looks into the recent advancements in secure multi-party computation (SMPC) for machine learning, a pivotal technology championing data privacy. We evaluate these applications from various aspects, including security models, requirements, system types, and service models, aligning with the IEEE's recommended practices for SMPC. Broadly, SMPC systems are divided into two categories: homomorphic-based systems, which facilitate computations on encrypted data, ensuring data remains confidential, and secret sharing-based systems, which disseminate data across parties in fragmented shares.

Oleksandr Lytvyn (2023) Machine learning methods require massive data collection to produce accurate predictions, raising privacy concerns. Secure Multi-Party Computation (SMPC) is one of the possible techniques to preserve data privacy during the computation process. Despite recent advances, the efficiency and scalability of SMPC in combination with machine learning remains an uncertain area. Moreover, various implementations of protocols exist that were optimized for the particular use cases, making the decision towards a suitable solution more complicated. The question of which protocol should be used under what circumstances naturally arises. In this case, when is defined as a security requirement and participant number. This work focuses on exploring the feasibility of using SMPC and machine learning in different settings.

Sergey Zapechnikov (2022) The study is devoted to the analysis of privacy-preserving machine learning (PPML) systems based on secure multi-party computations. It reviews PPML systems, analyses the goals and objectives of its application. A generalized model of PPML architecture is proposed, reflecting the main functional blocks of such systems. The formulation of the problem of secure multi-party computation is considered. The descriptions of cryptographic primitives and protocols used to implement multi-party secure computation protocols, including garbled circuits, secret sharing schemes, and homomorphic encryption are given. Systems secure against semi-honest and active adversaries are considered, both

based on universal modules for secure multi-party computations, and specialized ones designed to ensure the privacy of specific machine learning technologies, such as convolutional neural networks.

Chaochao Chen (2022) With the increasing demand for privacy protection, privacy-preserving machine learning has been drawing much attention from both academia and industry. However, most existing methods have their limitations in practical applications. On the one hand, although most cryptographic methods are provable secure, they bring heavy computation and communication. On the other hand, the security of many relatively efficient privacy-preserving techniques is being questioned, since they are non-provable secure. Inspired by previous works on privacy-preserving machine learning, we build a privacy-preserving machine learning framework by combining random permutation and arithmetic secret sharing via our compute-after-permutation technique. Our method is more efficient than existing cryptographic methods, since it can reduce the cost of element-wise function computation.

Harsh Chaudhari (2020) Privacy-preserving machine learning (PPML) via Secure Multi-party Computation (MPC) has gained momentum in the recent past. Assuming a minimal network of pair-wise private channels, we propose an efficient four-party PPML framework over rings $\mathbb{Z}2\ell$, FLASH, the first of its kind in the regime of PPML framework, that achieves the strongest security notion of Guaranteed Output Delivery (all parties obtain the output irrespective of adversary's behaviour). The state-of-the-art ML frameworks such as ABY3 by Mohassel et.al (ACM CCS'18) and SecureNN by Wagh et.al (PETS'19) operate in the setting of 3 parties with one malicious corruption but achieve the weaker security guarantee of abort. We demonstrate PPML with real-time efficiency, using the following custom-made tools that overcome the limitations of the aforementioned state-of-the-art– (a) dot product, which is independent of the vector size unlike the state-of-the-art ABY3, SecureNN and ASTRA by Chaudhari et.al (ACM CCSW'19), all of which have linear dependence on the vector size.

Privacy-preserving Tools for Machine Learning

The privacy-preserving tools for machine learning is the one around which all the contributions of this thesis are, secret-sharing-based secure multi-party computation. It is a cryptographic tool that allows multiple parties to jointly compute a function over their inputs while keeping those inputs private from each other. Formal definitions and details about secure multiparty computation are given in section 2.1 and in section 2.2 we give the basic protocols that are used to securely evaluate functions. Another privacy-preserving machine learning tool is homomorphic encryption, which is a form of encryption that allows computation to be done directly on encrypted data without needing to decrypt it first. This means that calculations can be carried out on the data in its encrypted form, and the result is also in an encrypted form. This allows for secure computation to occur without exposing the underlying data. This has many applications in the fields of cloud computing and data privacy, as it allows computations to be done without compromising the security of the data.

Importance of Privacy in Machine Learning

The concept of privacy has transformed over time, and in the field of machine learning, it encompasses more than simply being alone and undisturbed. Oxford defines privacy as "the state of being alone and not watched or disturbed by other people," but in the domain of machine learning, privacy incorporates not just the concerns of data owners, but additionally those of model owners and consumers. A data owner in the field of machine learning is a person or organisation that is ready to provide private information to help with the development of machine learning applications. Such information owners play a vital role in supplying the required data for algorithms. They have valid concerns regarding the privacy and security of their personal information. They desire confirmation that their private data will be managed and protected against unauthorised access or exploitation.

PPML based on secure two-party computations

The main motivation for increasing the number of participants in the SMPC protocols is the desire to achieve higher system performance and, if possible, provide users with stronger security guarantees for their data. Four-party computations make it possible to implement distributed computing procedures that are resistant to stronger adversary models than two-party and three-party ones. Protocols that provide security against an active adversary are of primary interest. Such security is provided by two key properties of the protocol, known as fairness and robustness. Let's give their informal definitions. Fairness is a guarantee that at the end of the protocol either all participants receive the same output data, or none of the participants receives any result, i.e. the protocol will be interrupted. There is a weakened definition of fairness, which allows the adversary to interrupt the participation of honest participants, but not allows the adversary to get any additional information. Robustness is a guarantee that honest participants will bring the protocol to the end and get the correct result when active adversary participates in protocol, even if she completely interrupts her participation in the protocol. One example of a four-party PPML system with advanced functionality that most fully implements the properties of security against an active adversary is Tetrad.

Real World Applications

Privacy-preserving tools in machine learning technologies could be applied in many sectors such as healthcare where medical records are highly sensitive, and preserving patient privacy is paramount. Privacy-preserving machine learning techniques like secure multi-party computation can be used to protect sensitive healthcare data from being accessed by unauthorized parties. By securely sharing and computing data from multiple parties, a secure protocol can be established to ensure that only the relevant parties have access to the data. Privacy-preserving tools can be used to analyse large sets of medical data, such as genomic data, without exposing individual identities or data points. Another area of application is law enforcement, where using machine learning to analyse large datasets in order to detect potential criminal activity is increasing. A concrete example of such an application was done by Roseman Labs, a startup that works on secure multi-party computation. They collaborated with the

NGO Sustainable Rescue to fight against human trafficking. Their solution uses secret-sharing to test if the name of a victim that sought an NGO for help is included in the law enforcement potential victims list, so that the law enforcement could act and find the traffickers. Keeping the victim's name secret would prevent them from going off-grid.

Privacy Preserving Machine Learning

Privacy-preserving machine learning corresponds to a collection of methods aimed at protecting the privacy of personal information while enabling the building and utilisation of models that use machine learning. It seeks to establish an equilibrium between the need for extracting valuable knowledge from data and the duty to protect the privacy rights of people in general. The significance of privacy-preserving machine learning begins with an increase in the number of technologies that utilise data and the rising worries regarding data privacy. As organisations acquire and analyse immense quantities of personal and sensitive data, protecting the privacy of those who use it becomes very important. There are numerous PPML techniques, each of which has its advantages and disadvantages. Some of the most prevalent techniques are:

Differential Privacy: A method that inserts noise in the data in a manner that respects the overall data distribution while making it harder to distinguish between individual records. **Homomorphic Encryption:** This encryption method enables calculations to be carried out using encrypted data without decryption. This in turn allows machine learning models to be trained on encrypted data with no disclosure of plaintext data.

Secure Multi-Party Computation: This method enables multiple parties to work alongside one another on a computation without disclosing each of their inputs. This can be utilised for training machine learning models upon data which is distributed throughout multiple parties without allowing any party access to the entire dataset.

RESEARCH METHODOLOGY

This study will utilise publicly accessible secondary data from an open-source repository. During collaborative machine learning, the dataset is partitioned horizontally or vertically into multiple subsets, with each subset representing a unique group of collaborators. To identify security and privacy threat during collaborative machine learning. To evaluate the efficiency and accuracy of the proposed techniques. Multiple open-source libraries will be used to help with the development and experimentation process. The objective is to simulate multiple participants sharing their data for collaborative machine learning, specifically multiple housing dataset owners would share their data to efficiently predict the housing price in Boston. Here each participant would train their model over the subset of the data to arrive at individual models and then. Each team's model is then combined to produce the final model. Experiments will be performed to assess various parameters, such as the number of data owners, different privacy budget values, and secure multi-party computation nodes, across a variety of hardware configurations. In addition, the incorporation of privacy-validating measures will facilitate an exhaustive assessment of the techniques' privacy-preserving capabilities. The results of these experiments will be beneficial to the progression of collaborative machine learning practises and influence the creation

of more reliable and privacy-conscious algorithms and methods. We will simulate multiple parties scenarios by splitting the Boston Housing Dataset into multiple groups, each holding a subset of data divided either horizontally or vertically.

RESULTS AND DISCUSSIONS

In order to set up a bloom filter correctly, you need to use hash functions that are totally independent and dispersed uniformly. In order to compare or index the data in the set, these functions allow for the assignment of any sort of data to an identifier.

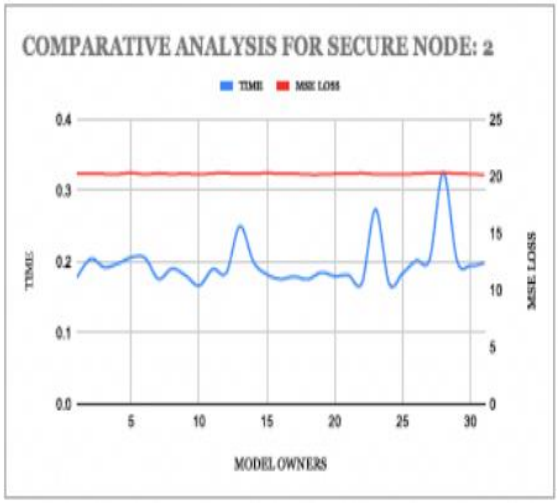
When people talk about hash functions, they usually mean the popular ones like SHA-256, MD5, or others like CRC32. Be cautious, nevertheless, when dealing with bloom filters. Although using several hash algorithms improves security, it also adds complexity and increases processing time. So, to get the most of their potential, it's crucial to choose the correct functions.

Table 1: The MSE Loss and Latency Results for Different Secure Nodes and Model Owners

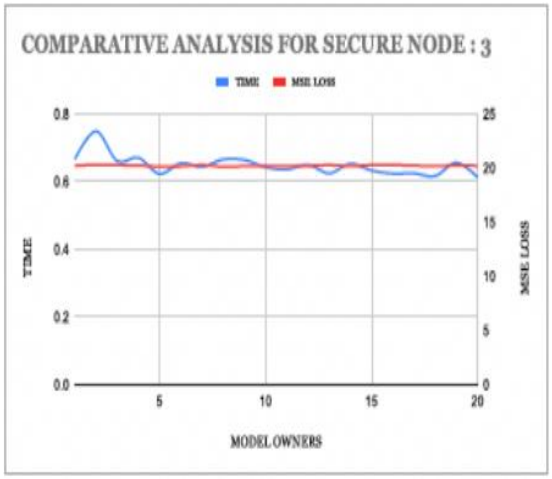
Secure Nodes	Models	Performance	Accurcay/ Loss
2	5	0.365245	21.3652
	10	0.235148	22.3854
	15	0.214875	24.2598
	20	0.698545	24.7412
3	5	0.634587	24.9874
	10	0.647852	24.5874
	15	0.656985	24.6985
	20	0.669874	24.6147
5	5	0.665245	30.5985
	10	0.678452	35.2695
	15	0.678966	45.2585
	20	0.688852	55.2698
15	5	0.654252	65.2554
	10	0.698742	75.2588
	15	0.635485	80.2266
	20	0.635874	81.2566
20	5	0.698755	90.2555
	10	0.698542	95.2156
	15	0.753321	95.8745
	20	0.786622	95.1234

Accuracy vs. Privacy

When the model parameters and inference values are sent across an SMPC cluster with two to ten secure nodes, the experimental findings reveal a loss of twenty to twenty-one percent in mean squared error (MSE) compared to the non-private MSE. This approach achieves its goal of protecting user privacy without compromising accuracy.

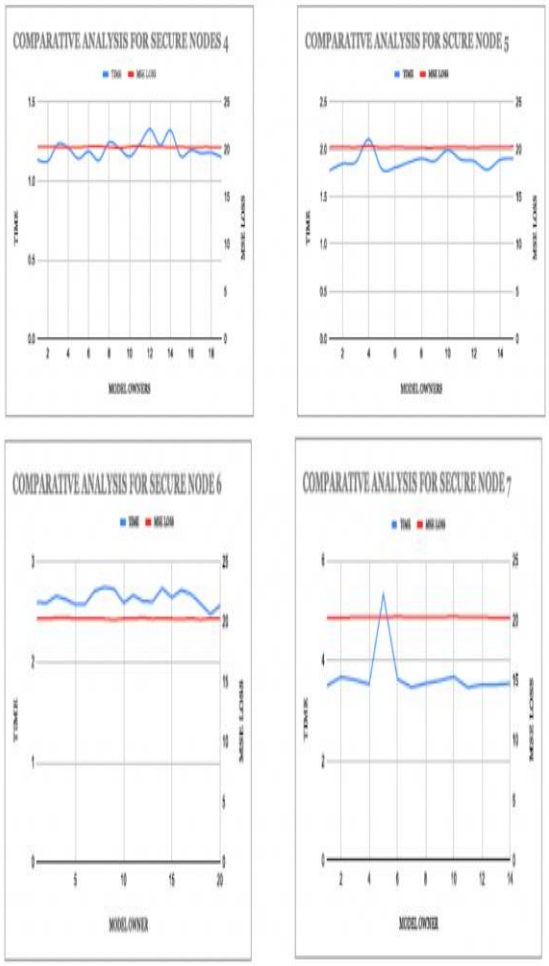


Graph: 1: Secure node vs. Performance vs. MSE Loss



Graph: 2: Secure node3 vs. Performance vs. MSE Loss

When it comes to privacy and security, cloud service providers and users have it bad. Customers relinquish ownership of their data throughout its transmission when they upload it to a public cloud server. Therefore, while storing data on the public cloud, it is important to carefully evaluate data privacy, availability, integrity, and trustworthiness.



Graph 3: Secure nodes 1-4 vs. Performance vs. MSE Loss

This study's results suggest two possible approaches: an OBA and an MSA. By using OBA, we want to mitigate security concerns related to public cloud storage. The first stage of the proposed approach is data security using OBA. Obtaining data with queries is the subsequent stage.

CONCLUSIONS

Collaborative machine learning plays a critical role in enhancing the performance and accuracy of machine learning systems by leveraging data contributed by multiple parties across domains. The proportion of packets lost on both servers increased as the quantity of malicious packets rose due to the additional labour needed to send notifications to disc. Existing privacy-preserving solutions have been unable to provide efficient and confidential ML training and inference, nor have they been able to strike a fine balance between multiple affecting parameters, such as accuracy, performance, and support for vertically partitioned datasets. To cope with these concerns and nurture collaboration, it is important to protect the sensitive information of machine learning stakeholders. By implementing comprehensive privacy-preserving techniques, such as Differential Privacy, Homomorphic Encryption, Secure Multi-Party Computation, and Federated Learning, it is possible to conduct collaborative machine learning without jeopardising privacy. Network intrusion detection systems that can manage gigabit rates without packet loss were developed using the architectural blueprints from the second phase. The network's capacity to detect intrusion attempts with any degree of accuracy was compromised when even the most sophisticated NIDS systems began to disregard packets over a certain threshold. The suggested solution included using a software load-balances to spread network traffic across many sensors, which improved Snort's speed and reliability over a fast network.

REFERENCES

1. *Sergey Zapechnikov (2022), "Secure multi-party computations for privacy-preserving*

machine learning", Procedia Computer Science, issn:1877-0509, vol.213, pages.523-527.https://doi.org/10.1016/j.procs.2022.11.100

2. *Massimo Piccardi (2024), "Secure Multi-Party Computation for Machine Learning: A Survey", IEEE Access, issn:2169-3536, vol.12, pages.1-1.DOI:10.1109/ACCESS.2024.3388992*
3. *Chaochao Chen (2022), "Towards secure and practical machine learning via secret sharing and random permutation", Knowledge-Based Systems, issn:0950-7051, vol.245(2), DOI:10.1016/j.knosys.2022.108609*
4. *Harsh Chaudhari (2020), "FLASH: Fast and Robust Framework for Privacy-preserving Machine Learning", Proceedings on Privacy Enhancing Technologies, issn:2299-0984, vol.2020(2), pages.459-480.DOI:10.2478/popets-2020-0036*
5. *Yang Zhou (2022), "From distributed machine learning to federated learning: a survey", Knowledge and Information Systems, issn:0219-1377, vol.64(4), pages.1-33.DOI:10. 1007/s10115-022-01664-x*
6. *Oleksandr Lytvyn (2023), "Efficiency and Security Trade-offs of Secure Multi-Party Computation for Machine Learning", Procedia Computer Science, issn:1877-0509, vol.225, pages.655-664.https://doi.org/10.1016/j.procs.2023.10.051*
7. *Farzad Tofigh (2024), "Secure Multi-Party Computation for Machine Learning: A Survey", IEEE Access, issn:2169-3536, doi:10.1109/ACCESS.2024.3388992*
8. *Venkatesh Kumar. M (2020), "An Efficient Secure Computation For Privacy Preserving Data Mining In Multi Party Computation (Mpc) – A Review", International Journal of Advanced Research in Engineering and Technology (IJARET), issn:0976-6499, vol.11, issue.10, pages.855-870.DOI: 10.34218/IJARET.11.10.2020.086*
9. *Reda Salamaa (2023), "Secure Multi-Party Computation for Collaborative Data Analysis", E3S Web of Conferences, issn:2267-1242, vol.399,https://doi.org/10.1051/e3sconf/202339904034*
10. *Hyunghoon Cho (2023), "Sequre: a high-performance framework for secure multiparty computation enables biomedical data sharing", Genome Biol., issn:1759-6653, vol.24, doi:10.1186/s13059-022-02841-5*