WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Review of Machine Learning and Cryptography Applications

Dr. Ganesh Kumar Dixit[1], Saurabh Kumar[2], Akshya Arya[3]

[1]Professor, AIDS, Arya College of Engineering, Kukas, Jaipur
[2]Student, AIDS, Arya College of Engineering, Kukas, Jaipur,
[3]Asst. Professor, AIDS, Arya College of Engineering, Kukas, Jaipur

ABSTRACT

Adversarial robust neural cryptography represents a fascinating convergence of two major fields: machine learning and cryptography. At its core, this approach involves training neural networks in the presence of an adversary—an opposing force that challenges the system and, in doing so, strengthens it. This adversary may itself be a neural network or a strategy shaped by one. Rather than being a threat, the adversary becomes a vital part of the learning process, helping refine the model's reliability and deepen its trustworthiness. In this evolving landscape, the relationship is mutual. Just as cryptographic thinking helps secure the workings of machine learning systems—shielding them from unauthorized access or manipulation—machine learning offers new ways to construct and test cryptographic tools, pushing the boundaries of what secure communication can look like. This paper offers a broad overview of this symbiotic relationship, charting how algorithms from each field are being used to enhance the other.

We examine how traditional cryptographic techniques are now solving modern machine learning challenges, and how deep learning models are transforming the design of cryptographic systems. A special focus is given to the growing and intriguing field of adversarial robustness, which explores how systems can become more resilient by embracing challenge and conflict as essential to their design.

Keywords: neural cryptography, deep learning, block ciphers, generative adversarial networks, adversarial robustness.

I. INTRODUCTION

Cryptography, at its heart, is the study of how we protect information—how we ensure that private data remains private, safe from those who might misuse it [1]. Traditionally, this field has relied on complex methods to disguise and later recover information, using techniques that are both difficult to reverse-engineer and carefully structured to allow authorized decoding [2]. A cornerstone of this discipline has been openness: cryptographic methods are often shared publicly so that experts around the world can examine them for flaws. This transparency, rather than undermining security, actually strengthens it. When vulnerabilities are discovered, the community can respond swiftly, adapting and improving the system [3]. This openness also aids the practical side of technology—enabling wider use, standardized implementation, and broader trust. Indeed, trust is a critical factor in cryptography. People are more likely to place confidence in systems they understand, or at least in those whose inner workings they know they could understand if they chose to [4]. Yet trust has long been a barrier to applying machine learning techniques within cryptography. Many early machine learning models operated as “black boxes,” providing little insight into how they made decisions. This lack of transparency slowed the integration of artificial intelligence into cryptographic systems. Recently, however, this has begun to change. The machine learning community has increasingly acknowledged the importance of trust and interpretability, giving rise to a growing field known as explainable AI [5–8]. These efforts are not just technical; they reflect a broader cultural shift toward openness and accountability in AI. This change has also encouraged researchers in other fields to explore machine learning's potential. In 2016, Abadi and colleagues [9] made waves by introducing a novel approach: using adversarial neural networks to create a model that could learn its own encryption strategy. Their work sparked widespread interest among both cryptographers and machine learning experts [10–13]. While their study garnered significant attention, other equally valuable contributions remain less well known. Understanding these developments in context is essential for appreciating the current landscape and the challenges that remain. This paper seeks to explore this evolving relationship between cryptography and machine learning. It begins with foundational concepts in deep learning and adversarial networks moves to machine learning algorithms inspired by cryptographic needs and then turns to cryptographic methods enriched by machine learning research. The paper concludes with a discussion of future directions and final reflections.

2.PRELIMINARIES

A. Deep Learning: Deep learning is a branch of machine learning that stands out for its depth—both literally and figuratively. These models are constructed from many layers, each filled with a multitude of parameters that help them learn complex patterns in data [14]. Most deep learning systems

take the form of artificial neural networks, inspired by the way biological brains process information. The architecture of these networks can vary: some layers may specialize in recognizing patterns in space (as in convolutional layers), others in patterns over time (as in recurrent layers), while some simply pass signals in a straightforward manner (as in fully-connected layers). Figure 1 offers a visual example of a deep learning model with a blend of such layers. Among these structures, convolutional neural networks (CNNs) have emerged as especially influential. They have proven remarkably adept at interpreting visual data—recognizing objects in photographs, identifying handwriting, and even helping medical professionals detect anomalies in scans [15]. Their rise has reshaped not only the field of computer science but also the way we interact with the visual world through technology.

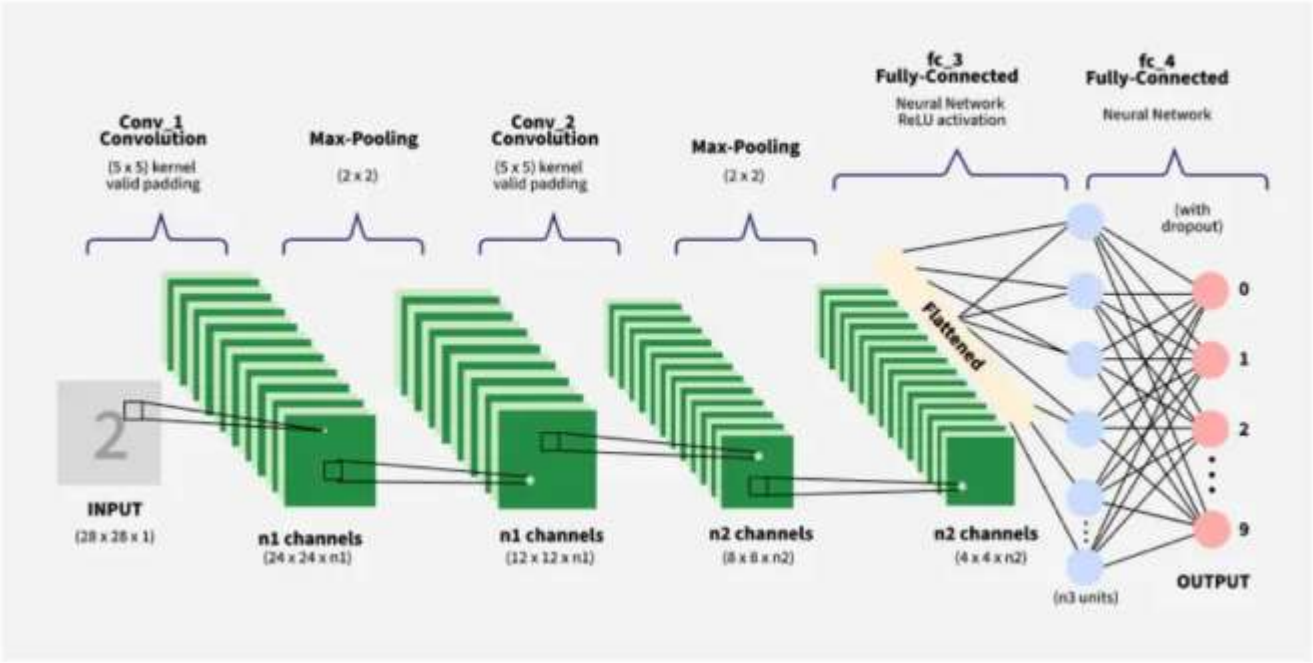

Figure: Convolutional Neural Network

B. Adversarial Neural Networks: In 2014, a groundbreaking idea was proposed by Ian Goodfellow and colleagues: Generative Adversarial Networks, or GANs [16], [17]. These networks introduced a novel concept—learning through competition. GANs are composed of two models locked in a dynamic interplay. One, the generator, tries to create data (often images) that mimic real examples. The other, the discriminator, attempts to detect whether the data it sees is real or a clever fake.back-and-forth resembles an artistic rivalry: the generator, like a forger, continuously improves its technique to fool the expert eye of the discriminator, which in turn becomes more discerning. As the generator learns to produce more convincing outputs and the discriminator becomes sharper in its judgment, the entire system advances in sophistication. This adversarial structure has opened up powerful possibilities—not just for generating realistic images, but also for exploring the very nature of learning, deception, and recognition in machines. GANs have since inspired a wave of research that pushes the boundaries of what artificial intelligence can create and understand.

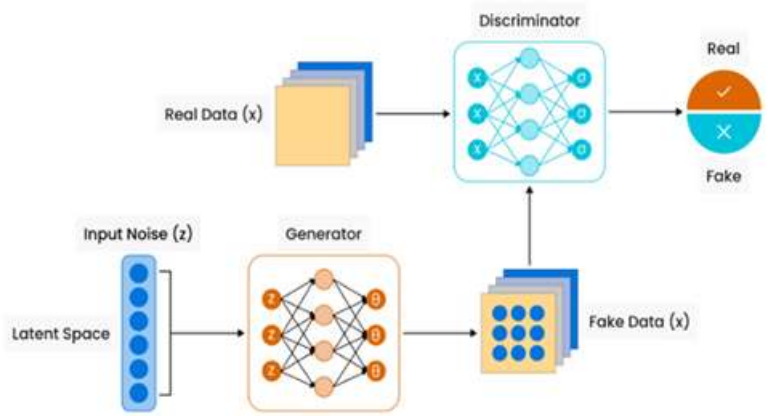

Figure: Adversarial Neural Networks

3. CRYPTOGRAPHY FOR MACHINE LEARNING

In many real-world situations, relying on a single source of data to train a deep learning model proves insufficient. Often, a richer, more accurate model requires input from multiple data holders—each with unique insights embedded in their private datasets. However, these data owners are understandably cautious. While they may be open to contributing to a shared model, they are not willing to expose their sensitive information. This tension between collaboration and confidentiality gives rise to what is known as collaborative deep learning.

In such scenarios, a central cloud-based system is often used to train the model, drawing from data provided by all participants. But since the data must remain private, specialized techniques are employed to ensure security. This paper focuses on two main approaches to collaborative learning: sharing encrypted training data and sharing encrypted gradients—both methods designed to protect the privacy of contributors during the learning process.

To uphold this privacy, the data—along with any computations performed in the cloud—must remain encrypted at all times. This is made possible through a technique known as fully homomorphic encryption (FHE), which allows computations to be performed on encrypted data without needing to first decrypt it [18].

A. Sharing Encrypted Training Data

Li and colleagues [19] proposed two innovative schemes—a basic version and a more advanced one—to enable secure, privacy-preserving collaboration among users training deep learning models in the cloud. In their approach, both the data owners and the cloud platform are considered honest-but-curious they follow the rules of the system faithfully but may still attempt to extract information if possible.

Their basic scheme relies on multi-key fully homomorphic encryption (MK-FHE) [20], which allows data encrypted with different keys to be processed jointly. Here's how it works:

1. Each data owner creates a unique set of public, private, and evaluation keys. They then encrypt their contributions—including training data, partial model weights, and target outputs—using their public key and send these encrypted components to the cloud.
2. Once the cloud has received the encrypted data from each participant, it begins the training process. Using the public and evaluation keys provided by the data owners, the cloud performs computations directly on the encrypted inputs—never needing to see the original, unencrypted data.
3. Finally, the data owners work together to decrypt the updated model parameters. They do this using a method known as secure multi-party computation (SMC) [21], which ensures that no single participant can see the full picture alone. Instead, through cooperation, they collectively recover their own individual updates without compromising the privacy of the others.

B. Sharing Encrypted Gradients

Another promising approach to collaborative deep learning involves sharing encrypted gradients—a method that focuses not on the raw data itself, but on the way that data shapes the learning process. In this context, gradients represent the adjustments a model makes during training, and even these, when exposed, can reveal sensitive information about the underlying data.

Phong and colleagues [29] built upon earlier work [30] that used a technique known as asynchronous stochastic gradient descent (ASGD) [31], [32] to allow different data owners to train their models independently and contribute selectively to a shared system. This earlier method—called gradient-selective ASGD—gave users the option to choose which learning signals (or gradients) they would share, offering a level of control over what parts of their data influenced the global model.

To further enhance privacy, the original approach also incorporated differential privacy, a method that introduces small amounts of random "noise" (via the Laplace mechanism [33]) to the shared gradients. The goal was to protect individuals' data even if gradients were intercepted. However, Phong and his team demonstrated that despite these efforts, both gradient-selective ASGD and differential privacy could still leak sensitive details. The subtle changes introduced weren't always enough to fully mask the underlying patterns. In response, they proposed a new method—one that relied on additively homomorphic encryption to protect the gradients more securely.

Their revised process works as follows:

1. Each data owner begins by downloading encrypted model weights from the cloud, using their own private.
2. Instead of sending this information in plain form, the data owner encrypts the scaled gradient (adjusted by a learning rate) using their secret key and transmits it back to the cloud.
3. Once the encrypted gradients are received, the cloud updates the global model—without ever decrypting the information. It simply performs addition operations on the encrypted values, as permitted by the homomorphic encryption scheme. In doing so, the model continues to learn and evolve, all while preserving the privacy of the individuals who contribute to it.

4. Conclusion

This paper has explored the emerging and dynamic relationship between cryptography and machine learning—two fields that, while distinct in origin, are increasingly shaping one another in profound ways. On one hand, cryptographic methods provide essential tools for safeguarding privacy and securing the data on which machine learning models depend. On the other, machine learning introduces novel techniques for modeling complexity, including the generation of pseudo-random, invertible functions that push the boundaries of what cryptographic systems can achieve.

One of the most compelling developments at this intersection is the focus on adversarial robustness—the ability to ensure that deep learning models, including those modeled after cryptographic systems like block ciphers, are resilient against manipulation. This work does more than strengthen security; it advances the broader goals of explainability and trustworthiness, qualities that are vital if AI is to be responsibly integrated into domains like healthcare, law, and science.

The growing field of neural cryptography offers a promising path forward. If supported by continued research and infrastructure investment, it could transform how we think about data protection. As the cost of data storage continues to fall, neural-based security models offer a scalable and efficient alternative for ensuring safe, private communication in a world where data is both abundant and vulnerable.

In this convergence of disciplines lies not just technical innovation, but a vision of collaboration that respects both intelligence and integrity—suggesting a future in which powerful learning systems do not compromise, but rather uphold, the fundamental values of privacy and trust.

5. Discussion and Future Directions

While homomorphic encryption has emerged as a powerful tool for protecting privacy in collaborative deep learning, its implementation is not the end of the story. Significant questions remain—particularly concerning the robustness of these systems in the face of adversarial threats. Adversarial attacks, in which carefully crafted inputs are used to mislead or destabilize a model, represent a persistent vulnerability. Researchers have proposed a range of defenses aiming to make deep learning systems more resistant to such attacks. Yet, true adversarial robustness remains an open frontier, requiring continued investigation and more refined approaches.

Equally important is the practical viability of adversarial neural cryptography. While the idea of deep learning models that can learn their own encryption mechanisms is intellectually exciting, the question of whether such models can be trusted and deployed in real-world security applications remains unresolved. Conventional (non-adversarial) deep learning systems are not inherently secure, and while adversarial models introduce new layers of complexity and adaptability, they currently lack the mathematical rigor and formal proofs that are traditionally required in cryptographic systems.

These challenges highlight a deeper tension at the heart of this field: the promise of innovation must be balanced by the demands of trust, verifiability, and ethical responsibility. Future work must therefore move beyond technical development alone. It must include a critical examination of the assumptions underlying machine learning in security contexts, along with collaborative efforts across disciplines to ensure that new technologies serve not just efficiency, but also integrity and accountability.

References

[1] B. Schneier, Applied Cryptography: Protocols, Algorithms, and Source Code in C, 2nd ed. Wiley, 1996.

[2] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, Handbook of Applied Cryptography. CRC Press, 1996.

[3] D. A. McGrew and D. R. Franklin, "The Importance of Open Standards in Cryptography," IEEE Security & Privacy, vol. 4, no. 5, pp. 87–90, 2006.

[4] R. Anderson, Security Engineering: A Guide to Building Dependable Distributed Systems, 2nd ed. Wiley, 2008.

[5] D. Gunning, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), 2017.

[6] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[7] Z. C. Lipton, "The mythos of model interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier," in Proc. of the 22nd ACM SIGKDD, pp. 1135–1144, 2016.

[9] M. Abadi, A. Chu, I. Goodfellow, H. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proc. of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318.

[10] M. Barni, K. Kallas, and T. Bianchi, "A privacy-preserving system for neural network training," in IEEE Transactions on Information Forensics and Security, vol. 11, no. 1, pp. 59–72, Jan. 2016.

[11] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in Proc. of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1310–1321, 2015.

[12] N. Papernot et al., "Semi-supervised knowledge transfer for deep learning from private training data," in Proc. of ICLR, 2017.

[13] R. Laine, P. Bourse, and M. Joye, "Fully homomorphic encryption over the integers for efficient privacy-preserving machine learning," in IEEE Transactions on Computers, vol. 68, no. 11, pp. 1652–1664, 2019.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems (NeurIPS), 2012.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems (NeurIPS), 2014.

[17] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," arXiv preprint arXiv:1701.00160, 2016.

[18] C. Gentry, "Fully homomorphic encryption using ideal lattices," in STOC '09: Proc. of the 41st annual ACM Symposium on Theory of Computing, pp. 169–178, 2009.

[19] W. Li, B. Chen, Y. Huang, and Y. Xiang, "Privacy-preserving federated brain tumour segmentation via multi-institutional cyclic training," in IEEE Transactions on Medical Imaging, vol. 40, no. 10, pp. 3054–3064, Oct. 2021.