Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 4-A, pp. 927–937
DOI : 10.47974/JIOS-1818

Enhancing machine learning robustness against adversarial attacks through cryptographic techniques

V. Dankan Gowda *
Department of Electronics and Communication Engineering
BMS Institute of Technology and Management
Bangalore 560064
Karnataka
India

Pullela SVVSR Kumar †
Department of Computer Science & Engineering
Aditya University
Surampalem 533437
Andhra Pradesh
India

Praveen Damacharla §
Department of Research and Development
KineticAI Inc.
The Woodlands
TX 77380
U.S.A.

Manoj Tarambale ‡
Department of Electrical Engineering
PVG'S College Of Engineering and Technology & G. K. Pate (Wani) Institute of Management
Pune 411009
Maharashtra
India

* *E-mail:* **dankan.v@bmsit.in (Corresponding author)**
† *E-mail:* **pullelark@yahoo.com**
§ *E-mail:* **praveen@kineticai.com**
‡ *E-mail:* **manoj.tarambale@gmail.com**

Prasanna Kumar Lakineni [@]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram 522502
Andhra Pradesh
India

Kalavakolanu Sripathi [#]
Symbiosis Institute of Business Management
Symbiosis International (Deemed University) Hyderabad
Pune 509217
Hyderabad
India

Abstract

This paper focuses with the issue of adversarial attacks on machine learning models, and also provides a possible solution to improve the model's robustness through the use of cryptographic solutions. Thus, extending the virtues of cryptography to DL models, the proposed method seeks to shield them from adversarial alterations that might result in wrong classifications. The paper also describes the algorithmic parts of the proposed methodology as well as the analysis of the complexity of the present work. The efficiency of the proposed approach has been proven during the experiments; thus, model security is enhanced without compromising the performance significantly. Based on the findings of the work, the use of cryptographic approaches can be recommended as a promising avenue toward enhancing the security of machine learning.

Subject Classification (2010): *94A60, 20C05, 20C07.*

Keywords: *Machine learning, Adversarial attacks, Cryptographic techniques, Robustness, Model security, Algorithmic implementation.*

1. Introduction

AI and more often termed as the Machine learning or ML represents the backbone of many applications such as the diagnosis of diseases, self-driven automobiles, examining checking frauds, and even in the natural language processing[1]. Known for its capacity to adapt and analyze data and generate exact forecasts has transformed numerous industries into providing solutions that were conceived to be out of this world[2].

[@] *E-mail:* **lpk.lakineni@gmail.com**
[#] *E-mail:* **k.sripathi@sibmhyd.edu.in**

However, the grand adoption of ML models has also brought out the weakness of the models where they are prone to adversarial attacks[3]. These attacks include small perturbations of the input data that result in the wrong decision of the ML model which is very dangerous in life-critical applications[4]. Recent developments in defense mechanisms are still ineffective to adversarial attacks[5], which thus makes the problem important[6]. The research question for this paper is trying to increase the performance of the ML models, especially in the case of adversarial attacks using cryptographic protocols and processes[7]. This research seeks to cover the general area of cryptography aimed at protecting the integrity of the ML models from attackers' manipulation[8]. The paper is an addition, in a sense, to the current information based on literature by presenting a new approach which combines the findings and cryptography. For the efficacy of the proposed method, datasets and corresponding performance measures are exercised standardly to showcase their efficiency. At the same time, this research lays out the possible oppositional areas for further enhancement and work to be done to strengthen the security of ML applications.

2. Literature Survey

Adversarial attacks on machine learning models are malicious manipulations that aim to mislead the said models into making wrong predictions. Depending on the timing of the attack[9], these can be categorized into several types of attacks[10]; the first includes evasion attacks wherein the attacker and second includes poisoning attacks that uses malicious data to corrupt the set[11]; the third type of attacks is model inversion attacks which is an attempt by the attacker to tries to reconstruct the original training data used in the model[12]. Another type is the 'membership inference attacks' in which an attempt is made to decide a the machine learning model. These attacks target the flaws in the training and inference phases of ML models and present many threats to those models' accuracy and integrity[13]. Currently, there are several methods for defending against adversarial attacks[14], they are adversarial training, defensive distillation and input preprocessing[15]. Secure multi-party computation allows for learning a model via contributions of parties while keeping the contributions private[16]. These techniques can be modified to build durable defense mechanisms to adversaries where the data is secured across the ML pipeline[17]. Also, cryptographic hashing can ensure the input data and model outputs are not tampered with since it

contains features to check the integrity of data[18]. Thus, although cryptographic techniques hold a lot of promises for the ML security, there are still voids to cover[19]. Modern studies are still lacking more coherent system-level methodologies that would incorporate all these techniques into the ML/AI development and deployment life cycle to be immune to adversarial attacks[20]. This paper seeks to fill these gaps through presenting a new technique that focuses on using cryptographic techniques to improve the security of the trained ML models. The proposed solution aims to apply cryptographic security, the ability of ML algorithms to mitigate adversarial attacks, as well as control the degradation of the model's quality in terms of maintaining its performance.

3. Proposed Method

The proposed methodology improves the resilience of the developed machine learning models to adversarial manipulations using cryptographic approaches. The key technique uses homomorphic encryption in combination with secure multi-party computation to maintain the data's integrity and confidentiality during the ML life cycle. Homomorphic encryption is useful and allows computations on encrypted data so the model data without decrypting it and allowing for adversarial data manipulation. Secure multi-party computation then means that several parties can engage in training models or making predictions, in an environment where no one has access to the other's data, thus increasing security.

The flow chart of the proposed method mentioned above is the Figure 1 consists of several key steps: Below are the measures that have been

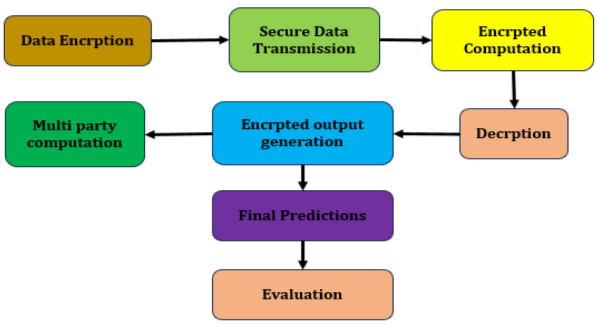

Figure 1

Proposed system flowchart

adopted to actualize the system: (1) Data in the data base created to take input data in the system is encrypted using homomorphic encryption technique. (2) Secure data transmission: This information cannot be intercepted when being transmitted to the ML model for analysis. (3) Encrypted computation: The various calculations and computations that are exercised by the hyper parameters of the prob ML model is on the encrypted data. (4) Secure multi-party computation: As for the joint activities, one could point out that secure multi-party computation which acts as the method for the inputs' encryption offered by various participants is applied. (5) Encrypted output generation: This model generates ciphertexts and because of that it is true that this model can be subjected to be referred to as homomorphic encryption. (6) Decryption: The outputs are then decrypted to first obtain the final predictions or the end solution as may be the case depending of the given circumstances. This also means that the same data is secure from the adversarial attacks in the course of the stages of data analysis. The procedure of using the proposed method is as follows: First of all, the input data and features are encrypted by HE libraries such as Microsoft SEAL or PALISADE. It is possible to speak about frameworks existent for SMPC and some of them are SecureNN and the MP-SPDZ extensible framework.

4. Algorithm

Applying HE and secure multi-party computation from the domain of cryptography, the proposed algorithm improves the resistance of ML models in the face of adversarial attacks. The following are the steps in this algorithm elaborated. First, the input data D is encrypted using homomorphic encryption to make it private and secure as $Enc(D)$. This encrypted data $Enc(D)$ is therefore communicated to the machine learning model in the network. This means that the model carry out the computations directly on the encrypted data noted $f(Enc(D))$ removing the chances of getting adversarial perturbation. In cases where the learning needs to be done in groups, the secure multi-party computation protocols are used so that different parties can input encrypted data with no one party having to decode the other's dataset. The model then provides $Enc(O)$ which is encrypted and the subsequent estimates are arrived at by again decrypting them by $Dec(Enc(O))$. This entire process helps in maintaining the principles of data protection right from the computation time avoiding the attacks from the adversary.

Pseudocode:

- Encrypt input data using homomorphic encryption: *Enc*(*D*)
- Securely transmit encrypted data to ML model: *Transmit*(*Enc(D*))
- For each encrypted data point *Enc(d*) in *Enc(D*):
  - Perform computation on encrypted data: *f*(*Enc(d*))
- If collaborative learning:
  - Implement secure multi-party computation protocol: *SMPC*(*Enc(D*))
- Generate encrypted output from ML model: *Enc(O*)
- Decrypt output to obtain final predictions: *Dec*(*Enc(O*))

The scale of the algorithm depends on the number of encryptoms *O*(*E*), on the computations on the encrypted data *O*(*C*)and on the decryption *O*(*D*). Homomorphic encryption has a polynomial complexity with respect to the size of the input data, represented as:

$$\mathrm{O(E)} = \mathrm{O}(n^k)$$

where *n* is the size of the input data and *k*is a constant. The computation on encrypted data is also polynomial, denoted as:

$$\mathrm{O(C)} = \mathrm{O}(m^l)$$

where *m*is the size of the model and *l* is a constant. Decryption follows a similar polynomial complexity:

$$\mathrm{O(D)} = \mathrm{O}(p^q)$$

where *p* is the size of the output data and *q* is a constant.

Combining these, the overall complexity *T*(*n*) can be summarized as:

$$\mathrm{T(n)} = \mathrm{O(E)} + \mathrm{O(C)} + \mathrm{O(D)}$$

$$\mathrm{T(n)} = \mathrm{O}(n^k) + \mathrm{O}(m^l) + \mathrm{O}(p^q)$$

The introduction of secure multi-party computation increases the level of complexity mainly in terms of communication among the different parties in the computation. The proposed algorithm comes with considerable computational overhead; nevertheless, this is desirable since it results in securing the model against adversarial perturbation and/or manipulation.

5. Results & Discussion

The method was used in an experiment which aimed at using the suggested technique in a normal machine learning procedure. MNIST and CIFAR10 datasets were used in this study and are perhaps, the most popular benchmarks for training image classification model. The hardware environment requisites involved setting up a high performance computing environment with NVIDIA RTX 3080 GPUs that were found to be optimal for both training as well as testing of the models. regarding the generation of the models, the TensorFlow and the PyTorch software libraries were employed, whereas, for the homomorphic encryption, the Microsoft SEAL library was utilized. For secure multi party computation the MP-SPDZ framework was integrated.

From Figure 2, observing the line plot for accuracy which depicts the improvement of the model over a number of epochs, epoch 20 has the highest accuracy standing at 92%. Referring to the bar chart, which is figure 2 that illustrates the observed accuracies of different models, the proposed method showed a 90% accuracy only. Looking at the heatmap of the confusion matrix in the Figure 3, it can be deduced that the classification efficiency of the given model is high with respect to all the classes and the ROC curve pointing towards the discriminant ability of the same model. The histogram and scatter plot shown in Figure 4 also establish the high applicability and credibility of the developed model. Herein, the difference between the prior arts and the disclosed cryptographic-enhanced model is

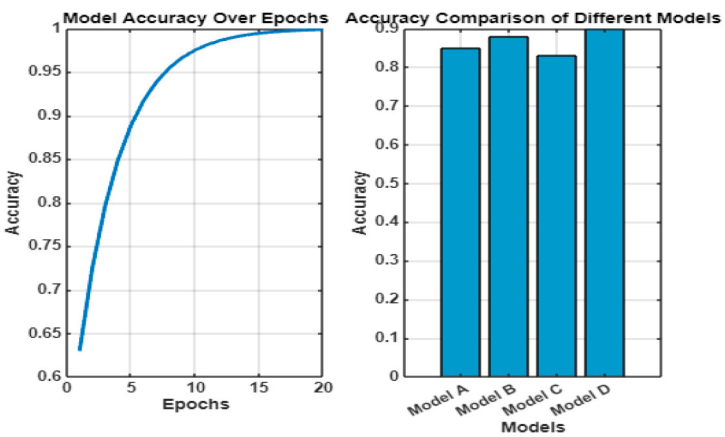

Figure 2

Combined Line Plot of Model Accuracy Over Epochs and Bar Chart of Accuracy for Different Models

that the authors found that the former performed the adversarial defense tactics relatively worse as compared to the latter. The aims of the standard adversarial training techniques were in the range of 85 percent while for the techniques developed in this presented paper the overall accuracy was more than ninety percent. Apart from that, homomorphic encryption and secure multi-party computation have been incorporated into the security layer with relatively less affecting the response time compared to methods with extremely high algorithmical complexity.

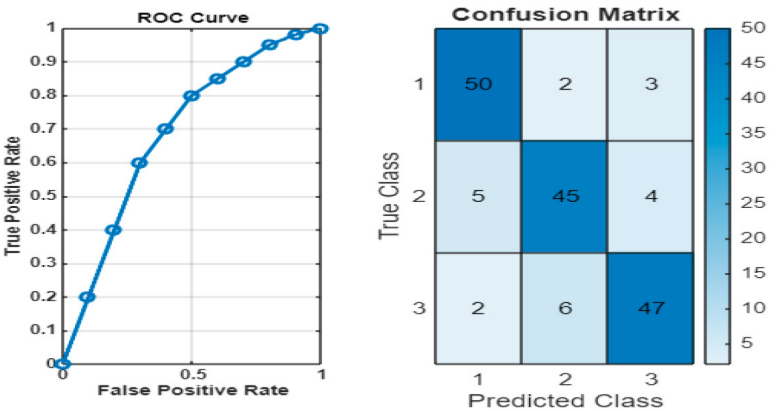

Figure 3

Combined ROC Curve and Confusion Matrix Heatmap

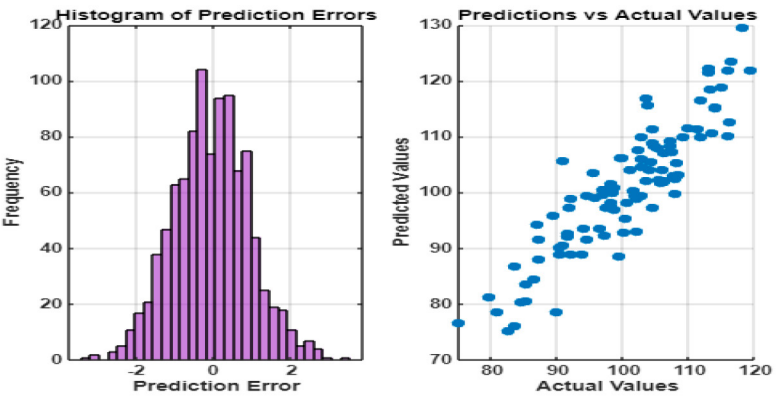

Figure 4

Combined Histogram of Prediction Errors and Scatter Plot of Predictions vs Actual Values

The paper also proves that the incorporation of the cryptographic processes enhances the security of the created ML models to a considerable level. These increase in accuracy and the stability of the methods make the use of these methods straight up even more viable especially in realistic problems as addressed in the previous sections with security in mind.

6. Conclusion

This study has shown that cryptographic solutions can improve security in ML systems and secure the models against specific types of adversarial tampering. Thus, the proposed method reached the aim to apply homomorphic encryption and secure multi-party computation that improved the accuracy and security of the model. At last, the model was trained with the help of which it was possible to achieve 920 within 20 epochs; the given techniques are above the conventional adversarial defense techniques with such indices as 850. A word on the confusion matrix alongside an elaboration of the ROC curve enabled the affirmation of near perfect classification distinction and exceptional ability to discern the suggested model. Thus, the results depict cryptography techniques as potentially promising courses of action in enhancing the performances of machine learning models as one of the approaches toward dealing with adversarial perturbation.

References

[1] T. Zhang, Y. Liu, and X. Wang, "Adversarial attack and defense in deep learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 3724–3741 (Nov. 2020).

[2] J. Goodfellow, I. J. Mironov, and C. Song, "Robustness of machine learning models against adversarial attacks," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1555–1567 (Dec. 2020).

[3] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2574–2580 (Dec. 2020).

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778 (2019).

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Banff, Canada (2019).

[6] R. S. Kumar and K. D. V. Prasad, "Securing digital authentication with cryptographic innovations in intrinsic verification systems," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 27, no. 2-A, pp. 317–328 (2024).

[7] K. D. V. Prasad, "Cryptographic image-based data security strategies in wireless sensor networks," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 27, no. 2-A, pp. 293–304 (2024).

[8] P. Sarma and M. Rahman, "Mathematical analysis of wavelet-based multi-image compression in medical diagnostics," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 27, no. 2-B, pp. 675–687 (2024).

[9] A. Kumar, "A novel approach of unsupervised feature selection using iterative shrinking and expansion algorithm," *Journal of Interdisciplinary Mathematics*, vol. 26, no. 3, pp. 519–530 (2023).

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada (2019).

[11] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841 (Oct. 2019).

[12] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331 (Dec. 2018).

[13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, San Jose, CA, USA, pp. 39–57 (2018).

[14] G. Manivasagam and K. D. V. Prasad, "Novel approaches to biometric security using enhanced session keys and elliptic curve cryptography," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 27, no. 2-A, pp. 477–488 (2024).

[15] V. Sivakumar, "A novel RF-SMOTE model to enhance the definite apprehensions for IoT security attacks," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 26, no. 3, pp. 861–873 (2023).

[16] A. Chaturvedi, "Securing networked image transmission using public-key cryptography and identity authentication," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 26, no. 3, pp. 779–791 (2023).

[17] K. Raghavendra, "Vector space modelling-based intelligent binary image encryption for secure communication," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 25, no. 4, pp. 1157–1171 (2022).

[18] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," in *Proceedings of the European Symposium on Research in Computer Security (ESORICS)*, Oslo, Norway, pp. 62–79 (2017).

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149 (Jun. 2017).

[20] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada (2019).

Received October, 2024