Robust and Fair AI: Addressing Adversarial Attacks, Data Bias and Privacy Challenges

Samuel Verghese (3120522) *

August 5, 2024

Abstract

This paper explores ways to make AI systems more reliable and secure by looking at three main areas: handling attacks, reducing bias, and protecting privacy. We test how well AI models work on normal and tricky data, comparing a basic model with one trained to handle attacks better. We also check how different methods to fix biased data affect the model's performance. Finally, we examine how adding privacy protections can affect the model's accuracy. All these tests were done using Python, helping us understand how to build AI that is strong against attacks, fair, and keeps user data safe.

1 Introduction

As AI becomes a bigger part of our lives, it's important to make sure these systems are both reliable and safe. AI models face a few big challenges: they can be tricked by special attacks, they might make biased decisions based on flawed data, and there are concerns about keeping user information private.

First, AI models can be fooled by adversarial attacks, where inputs are deliberately altered to confuse the system. This can hurt the model's performance on normal data and make it less trustworthy. We need to understand how these attacks work and find ways to train models to handle them better.

Second, if the data used to train AI models is imbalanced, the models can be biased and unfair. Techniques like oversampling (adding more of the minority class) or under sampling (reducing the majority class) can help improve fairness. It's important to test these methods to see how well they work.

Lastly, differential privacy helps protect user data, but it can sometimes lower the accuracy of the AI model. We need to find a balance between keeping user information safe and ensuring the model works well.

2 Literature Review

Adversarial training is a method used to make machine learning (ML) models more robust against attacks where inputs are intentionally altered to confuse the model (Gungor et al., 2022; Lin et al., 2022). By including these tricky examples during training, the model learns to handle such manipulations better. However, this approach isn't perfect. It might reduce the model's performance on normal, clean data and may not defend against every type of adversarial attack (Jha et al., 2018). Additionally, models trained with adversarial examples might become vulnerable to other types of attacks, such as those that try to infer private information from the model's behavior (Mittal et al., 2019). Thus, while adversarial training can help, it has limitations and may

*MSc Computer Science. SRH Hochschule, Berlin

introduce new issues that need to be addressed (Gungor et al., 2022; Jha et al., 2018; Lin et al., 2022; Mittal et al., 2019).

Mitigating bias in datasets is crucial for fair and accurate machine learning models. When datasets are imbalanced—meaning some classes are overrepresented compared to others—it can lead to biased model outcomes (Imani & Arabnia, 2023; Vargas et al., 2022). Common methods to address this include oversampling the less common class, undersampling the more common class, or using techniques like the Synthetic Minority Over-sampling Technique (SMOTE) (Imani & Arabnia, 2023). Other strategies, such as spatial filtering and weighted sampling, are used to reduce bias in specific contexts like species distribution modeling and geological data (Gutierrez-Velez & Wiese, 2020; Liu et al., 2020). These methods show that while resampling is useful, other techniques may also be needed to handle different types of biases effectively (Gutierrez-Velez & Wiese, 2020; Liu et al., 2020). Continued research is essential to improve these methods.

Differential privacy is a technique used to protect individual data in machine learning models by ensuring that the inclusion or exclusion of a single person's data does not significantly affect the model's output (Ting, 2022). While it's well-studied in traditional ML, its use in quantum machine learning (QML) is emerging. Recent studies show that differential privacy can be applied to QML without greatly affecting accuracy, and combining it with quantum federated learning can enhance security and efficiency (Watkins et al., 2023; Tseng et al., 2023). However, there is a trade-off between privacy and accuracy. For example, adding differential privacy to fingerprint recognition systems can reduce accuracy as privacy measures become stricter (Mohammadi et al., 2023). Similarly, deep differential privacy models can balance privacy and accuracy by adjusting privacy parameters (Liu & Arif, 2022; Phan & Tran, 2023). Overall, differential privacy remains a crucial method for protecting user data while maintaining model performance, with ongoing research aimed at improving this balance (Liu & Arif, 2022; Mohammadi et al., 2023; Phan & Tran, 2023; Ting, 2022; Tseng et al., 2023; Watkins et al., 2023).

3 Methodology

All experimental code and scripts used in this study are available in the GitHub repository at `https://github.com/Spinal-Tap369/Robustness-in-ML`. This repository contains the complete implementation of the methodologies described in this paper.

3.1 Defense against Adversarial Attacks

Adversarial attacks in machine learning are when someone changes the input data on purpose to trick the model into making wrong predictions. These changes are usually small and hard for people to notice but can make the model's results very different (Mani et al., 2019; Oh et al., 2022). One common method for generating adversarial examples is the Fast Gradient Sign Method (FGSM), which adds perturbations to the input data based on the gradient of the loss function with respect to the input (Liu et al., 2019).

Adversarial training is a method to make models stronger against attacks by using tricky examples during training. It mixes both normal and altered examples so the model can learn to handle these tricky situations better (Mani et al., 2019; Pang et al., 2020).

For the purposes of this research, a Convolutional Neural Network (CNN) was used for image classification, featuring several convolutional layers, batch normalization, ReLU activations, max pooling, and dropout. The model is trained on the CIFAR-10 dataset, which includes 60,000 images across 10 categories. The images are preprocessed with normalization.

The training process starts with clean data

using cross-entropy loss and the Adam optimizer. After training, we evaluate the model's susceptibility to adversarial examples by applying FGSM. Subsequently, the model is retrained with a mix of clean and adversarial examples to enhance its robustness. This involves 10 additional epochs of training with a loss function that combines clean and adversarial losses, adjusted by a balance factor. A learning rate scheduler is also used to fine-tune the training.

Finally, the model's accuracy is evaluated on both clean and adversarial test datasets to assess the effectiveness of the adversarial training.

3.2 Bias Mitigation

Biased datasets have an imbalance in the class distribution, where one class significantly outnumbers the other class (Alkhateeb & Maolood, 2019; Peng & Wang, 2022). For the purposes of this paper, the fradulent credit card transactions dataset was used which is highly biased towards non-fraudulent credit card transactions.

The dataset was then split into training and test sets using the train-test split, with 70% for training and 30% for testing.

An XGBoost classifier was defined and trained on the original imbalanced dataset. The model's performance was evaluated using a classification report and confusion matrix. To mitigate the class imbalance, SMOTE was used for oversampling the minority class, and RandomUnderSampler was used for undersampling the majority class. Two pipelines were created: one for oversampling and one for undersampling, both including the XGBoost model.

Each pipeline was fitted to the training data and evaluated on the test data. Classification reports and confusion matrices were generated to compare the performance of the model on the oversampled and undersampled datasets. This approach helped assess the effectiveness of different sampling techniques in mitigating bias and improving model performance.

3.3 Differential Privacy

Differential privacy is a technique used to protect individual data in machine learning models by ensuring that the inclusion or exclusion of a single person's data does not significantly affect the model's output (Ting, 2022).

For this paper, differential privacy was performed on the Iris dataset.

A standard logistic regression model without differential privacy is trained on the standardized training data. After training, the model's accuracy is evaluated on the test set using accuracy score. This step establishes a baseline accuracy for comparison with differentially private models.

In differential privacy, the epsilon parameter, denoted as ϵ, gauges the level of privacy protection. Smaller ϵ values indicates stronger privacy guarantees but potentially lower accuracy of the output data (Adewole et al., 2019; Nanayakkara et al., 2023).

Next, a differentially private logistic regression model from the diffprivlib library is trained with various epsilon values and the accuracy scores are compared.

4 Results

4.1 Adversarial Defense Training

The initial model, tested on clean data, achieved an accuracy of 81.11%. However, its performance dropped significantly to 22.38% when evaluated with adversarial examples, indicating vulnerability to such attacks. After implementing adversarial training over 10 epochs, the model's accuracy improved. Specifically, during training, the model's accuracy on clean data slightly decreased from 81.11% to 77.98%. In contrast, its accuracy on adversarial data increased from 22.38% to 47.61%. This demonstrates that adversarial training effectively enhances the model's robustness against adversarial attacks while maintaining a reasonably high accuracy on clean data.

4.2 Bias Mitigation

For the credit card fraud detection dataset, the original imbalanced dataset showed very high accuracy at 100%, but this masks poor performance on the minority class (fraudulent transactions), with a recall of only 0.82. The confusion matrix revealed that the model was highly effective at detecting non-fraudulent transactions but struggled with fraudulent ones.

After applying oversampling, which balances the number of fraudulent and non-fraudulent transactions, the model's recall for fraudulent transactions improved to 0.87, though the precision dropped to 0.73. This approach led to an overall accuracy of 100%. The confusion matrix indicated a more balanced detection between the two classes.

Conversely, undersampling the majority class resulted in a model that had high recall for fraudulent transactions (0.93) but very low precision (0.04). This method reduced the overall accuracy to 96% and highlighted a trade-off where the model becomes better at detecting fraud but at the cost of increased false positives.

4.3 Differential Privacy

The logistic regression model without privacy constraints achieved a perfect accuracy of 100%. When applying differential privacy with an epsilon value of 10, the model's accuracy dropped significantly to 42%, indicating that high privacy levels reduce model performance. With a higher epsilon value of 60, which offers less privacy protection, the accuracy improved to 84%. This shows a clear trade-off between privacy and model accuracy, with higher epsilon values allowing the model to perform better while still providing some level of privacy.

5 Future Works

Future research should focus on several key areas to enhance the methods discussed in this study.

To improve adversarial training, further studies should explore techniques to minimize the trade-off between robustness and accuracy on clean data. Researchers could investigate advanced adversarial training methods, such as combining different types of adversarial examples or using more sophisticated loss functions, to achieve better performance without compromising too much on accuracy.

Exploring hybrid approaches that combine multiple strategies or developing new methods tailored to specific types of data imbalances could offer more balanced solutions for Bias Mitigation strategies.

Research into differential privacy should aim to refine the balance between privacy and accuracy. Investigating methods to adjust the privacy parameter (epsilon) based on the data and model requirements could help improve performance while maintaining strong privacy protections.

6 Conclusion

This study explored methods to improve machine learning models by addressing three key areas: adversarial attacks, bias mitigation, and differential privacy. Adversarial training showed promise in making models more resilient to attacks, although it slightly reduced performance on clean data. Bias mitigation techniques, like oversampling and undersampling, helped balance model performance across different classes but highlighted trade-offs between precision and recall. Differential privacy demonstrated that higher privacy levels can lower model accuracy, but balancing privacy and performance is crucial. Overall, these methods are essential for developing more reliable, fair, and secure machine learning systems. Future work will focus on refining these approaches to enhance their effectiveness and applicability.

7 References

1. Liu, Y., Mei, X., Yang, T., Mao, S., & Zhao, X. (2019). Sensitivity of Adversarial Perturbation in Fast Gradient Sign Method. *Institute of Electrical Electronics Engineers*. https://doi.org/10.1109/ssci44817.2019.9002856

2. Mani, N., Moh, M., & Moh, T.-S. (2019). Towards Robust Ensemble Defense Against Adversarial Examples Attack. *Institute of Electrical Electronics Engineers*. https://doi.org/10.1109/globecom38437.2019.9013408

3. Oh, C., Xompero, A., & Cavallaro, A. (2022). Chapter 15 - Visual adversarial attacks and defenses. In *Advanced Methods and Deep Learning in Computer Vision* (pp. 511–543). Elsevier. https://doi.org/10.1016/b978-0-12-822109-9.00024-2

4. Pang, N., Ji, Y., Pan, Y., & Hong, S. (2020). Efficient Defense Against Adversarial Attacks and Security Evaluation of Deep Learning System (pp. 592–602). *Springer*. https://doi.org/10.1007/978-3-030-62460-6_53

5. Du, X., Tian, Z., Guizani, M., & Susilo, W. (2021). Introduction to the Special Section on Artificial Intelligence Security: Adversarial Attack and Defense. *IEEE Transactions on Network Science and Engineering, 8(2)*, 905–907. https://doi.org/10.1109/tnse.2021.3073637

6. Lin, Y.-D., Pratama, J.-H., Sudyana, D., Lai, Y.-C., Hwang, R.-H., Lin, P.-C., Lin, H.-Y., Lee, W.-B., & Chiang, C.-K. (2022). ELAT: Ensemble Learning with Adversarial Training in defending against evaded intrusions. *Journal of Information Security and Applications, 71*, 103348. https://doi.org/10.1016/j.jisa.2022.103348

7. Gungor, O., Rosing, T., & Aksanli, B. (2022). DENSE-DEFENSE: Diversity Promoting Ensemble Adversarial Training Towards Effective Defense. https://doi.org/10.1109/sensors52175.2022.9967204

8. Alkhateeb, Z. K., & Maolood, A. T. (2019). Machine Learning-Based Detection of Credit Card Fraud: A Comparative Study. *American Journal of Engineering and Applied Sciences*. https://doi.org/10.3844/ajeassp.2019.535.542

9. Jha, S., Jha, S., Jalaian, B., & Jang, U. (2018). Detecting Adversarial Examples Using Data Manifolds. *Institute of Electrical and Electronics Engineers*. https://doi.org/10.1109/milcom.2018.8599691

10. Peng, H., & Wang, J. (2022). Unbalanced Data Processing and Machine Learning in Credit Card Fraud Detection. *Research Square Platform LLC*. https://doi.org/10.21203/rs.3.rs-2004320/v1

11. Mittal, P., Shokri, R., & Song, L. (2019). Privacy Risks of Securing Machine Learning Models against Adversarial Examples. https://doi.org/10.48550/arxiv.1905.10291

12. Adewole, A., Olayiwola, O., & Udeh, S. (2019). Differential Privacy: A Non-Stochastic Approach to Privacy Parameter Generation. *Journal of Research and Review in Science*, 6(1). https://doi.org/10.36108/jrrslasu/9102/60(0190)

13. Nanayakkara, P., Kaptchuk, G., Smart, M., Cummings, R., & Redmiles, E. (2023). What Are the Chances? Explaining the Epsilon Parameter in

Differential Privacy. *Cornell University*. https://doi.org/10.48550/arxiv.2303.00738

14. Gutierrez-Velez, V. H., & Wiese, D. (2020). Sampling bias mitigation for species occurrence modeling using machine learning methods. *Ecological Informatics, 58*, 101091. https://doi.org/10.1016/j.ecoinf.2020.101091

15. Liu, W., Sivila, L., Pyrcz, M. J., Scott Hamlin, H., & Ikonnikova, S. (2020). Demonstration and Mitigation of Spatial Sampling Bias for Machine-Learning Predictions. *SPE Formation Evaluation, 24*(01), 262–274. https://doi.org/10.2118/203838-pa

16. Werner De Vargas, V., Victória Barbosa, J. L., Dos Santos Costa, R., Schneider Aranda, J. A., & Da Silva Pereira, P. R. (2022). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems, 65*(1), 31–57. https://doi.org/10.1007/s10115-022-01772-8

17. Imani, M., & Arabnia, H. R. (2023). Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis. *MDPI AG*. https://doi.org/10.20944/preprints202308.1478.v1

18. Ting, C.-K. (2022). Quest for the Balance of AI and Privacy [Editor's Remarks]. *IEEE Computational Intelligence Magazine, 17*(3), 2. https://doi.org/10.1109/mci.2022.3180649

19. Watkins, W. M., Chen, S. Y.-C., & Yoo, S. (2023). Quantum machine learning with differential privacy. *Scientific Reports, 13*(1). https://doi.org/10.1038/s41598-022-24082-z

20. Tseng, H.-H., Rofougaran, R., Chen, S., & Yoo, S. (2023). Federated Quantum Machine Learning with Differential Privacy. https://doi.org/10.48550/arxiv.2310.06973

21. Liu, Q., & Arif, M. (2022). Privacy Protection Technology Based on Machine Learning and Intelligent Data Recognition. *Security and Communication Networks, 2022*, 1–9. https://doi.org/10.1155/2022/1598826

22. Phan, T., & Tran, H. (2023). Consideration of Data Security and Privacy Using Machine Learning Techniques. *International Journal of Data Informatics and Intelligent Computing, 2*(4), 20–32. https://doi.org/10.59461/ijdiic.v2i4.90

23. Mohammadi, M., Sabry, F., Malluhi, Q., & Labda, W. (2023). Privacy-preserving Deep-learning Models for Fingerprint Data using Differential Privacy. *Association for Computing Machinery*. https://doi.org/10.1145/3579987.3586568