REFERENCES

[1] [n. d.]. The Project GitHub Repo. https://github.com/ykumar2020/datapoisononing. Last accessed: July 6, 2024.

[2] Arif Agustyawan. 2020. Fish Dataset. https://www.kaggle.com/datasets/arifagustyawan/fresh-and-not-fresh-fish-dataset/data. Last accessed: July 5, 2024.

[3] David P. Benalcazar, Juan E. Zambrano, Daniel Bastias, Carlos A. Pérez, and Kevin Bowyer. 2020. A 3D Iris Scanner From a Single Image Using Convolutional Neural Networks. *IEEE Access* 8 (2020), 98584–98599.

[4] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389* (2012). https://arxiv.org/abs/1206.6389 Accessed: December 29, 2024.

[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1712.04248

[6] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*. IEEE, 39–57. https://arxiv.org/abs/1608.04644

[7] Alice E Cinà, Kathrin Grosse, Ambra Demontis, Battista Biggio, Fabio Roli, Marcello Pelillo, et al. 2024. Machine learning security against data poisoning: Are we there yet? *Computer* 57, 3 (2024), 26–34.

[8] Hugging Face. 2024. Transformers Documentation. https://huggingface.co/docs/transformers/en/index. Accessed: October 3, 2024.

[9] Sina Farhadkhani, Rachid Guerraoui, Nitin Gupta, and Robin Pinot. 2024. On the Relevance of Byzantine Robust Optimization Against Data Poisoning. *arXiv preprint arXiv:2405.00491* (2024).

[10] Dejaun Gayle. 2025. AdverseraGuardV2 GithubRepo. https://github.com/DejaunG/AdverseraGuardV2. Accessed: 2025-07-03.

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014). https://arxiv.org/abs/1412.6572 Accessed: December 29, 2024.

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

[13] Brendan Hannon, Yulia Kumar, Dejaun Gayle, J. Jenny Li, and Patricia Morreale. 2024. Robust Testing of AI Language Model Resiliency with Novel Adversarial Prompts. *Electronics* 13 (2024), 842. https://doi.org/10.3390/electronics13050842

[14] Brendan Hannon, Yulia Kumar, Peter Sorial, J. Jenny Li, and Patricia Morreale. 2023. From Vulnerabilities to Improvements: A Deep Dive into Adversarial Testing of AI Models. In *Proceedings of the 21st International Conference on Software Engineering Research & Practice (SERP 2023).* Presented.

[15] H Kasyap and Sukumar Tripathy. 2024. Beyond data poisoning in federated learning. *Expert Systems with Applications* 235 (2024), 121192.

[16] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1885–1894.

[17] Yulia Kumar, Patricia Morreale, Peter Sorial, Justin Delgado, J. Jenny Li, and Pedro Martins. 2023. A Testing Framework for AI Linguistic Systems (testFAILS). In *Proceedings of the 2023 IEEE International Conference on Artificial Intelligence Testing (AITest).* https://doi.org/10.1109/AITest58265.2023.00004

[18] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. 2017. Trojaning attack on neural networks. *arXiv preprint arXiv:1707.07860* (2017). https://arxiv.org/abs/1707.07860

[19] Aleksander Madry. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017). https://arxiv.org/abs/1706.06083 Accessed: December 29, 2024.

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)* (2018).

[21] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 1765–1773.

[22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2574–2582. https://arxiv.org/abs/1511.04599

[23] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, et al. 2018. Ray: A distributed framework for emerging AI applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18).* USENIX Association, 561–577.

[24] OpenAI. 2022. Techniques for training large neural networks. https://openai.com/index/techniques-for-training-large-neural-networks/ Accessed: December 29, 2024.

[25] Martin Pawelczyk, J. Z. Di, Yifei Lu, Gautam Kamath, Aadirupa Sekhari, and Seth Neel. 2024. Machine unlearning fails to remove data poisoning attacks. *arXiv preprint arXiv:2406.17216* (2024).

[26] Anh Pham, Maria Potop-Butucaru, Sébastien Tixeuil, and Serge Fdida. 2024. Data Poisoning Attacks in Gossip Learning. In *International Conference on Advanced Information Networking and Applications*. Springer, 213–224.

[27] Alexander Sergeev and Mike Del Balso. 2018. Horovod: fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799* (2018).

[28] Jacob Steinhardt, Pang Wei W. Koh, and Percy S. Liang. 2017. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, Vol. 30.

[29] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* 23, 5 (2019), 828–841. https://doi.org/10.1109/TEVC.2019.2890858

[30] Christian Szegedy. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013). https://arxiv.org/abs/1312.6199 Accessed: December 29, 2024.

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[32] Oguzhan Ulucan, Derya Karakaya, and Mehmet Turkan. 2020. A Large-Scale Dataset for Fish Segmentation and Classification. https://www.kaggle.com/datasets/crowww/a-large-scale-fish-dataset. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, Istanbul, Turkey, 1–5. https://doi.org/10.1109/ASYU50717.2020.9259867 Last accessed: July 5, 2024.

[33] Weitian Wang and Soheil Feizi. 2024. Temporal robustness against data poisoning. In *Advances in Neural Information Processing Systems*, Vol. 36.

[34] Yulin Wang, Tongyu Sun, Shanshan Li, Xiong Yuan, Weiping Ni, Ekram Hossain, and H. Vincent Poor. 2023. Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey. *arXiv preprint arXiv:2303.06302* (2023).

[35] Chuan Xu, Xiaofang Zhu, Wenjing He, Yiran Lu, Xiang He, Ziyan Shang, Jianhua Wu, Kai Zhang, Yifan Zhang, Xiaoyi Rong, et al. 2019. Fully Deep Learning for Slit-Lamp Photo Based Nuclear Cataract Grading. In *International Conference on Medical Image Computing and Computer-Assisted Intervention.*

[36] Zhang Zheng, Zhe Li, Chun Huang, Shi Long, Ming Li, and Xuemin Shen. 2024. Data poisoning attacks and defenses to LDP-based privacy-preserving crowdsensing. *IEEE Transactions on Dependable and Secure Computing* (2024).