Supplementary: Analyzing Federated Learning through an Adversarial Lens

Arjun Nitin Bhagoji [1 *] **Supriyo Chakraborty** [2] **Prateek Mittal** [1] **Seraphin Calo** [2]

1. Further experimental details

1.1. Datasets

Fashion-MNIST: This dataset serves as a drop-in replacement for the commonly used MNIST dataset (LeCun et al., 1998), which is not representative of modern computer vision tasks. It consists of 28×28 grayscale images of clothing and footwear items and has 10 output classes. The training set contains 60,000 data samples while the test/validation set has 10,000 samples.

Adult Census: It has over 40,000 samples containing information about adults from the 1994 US Census. The classification problem is to determine if the income for a particular individual is greater (class '0') or less (class '1') than $\$50,000$ a year.

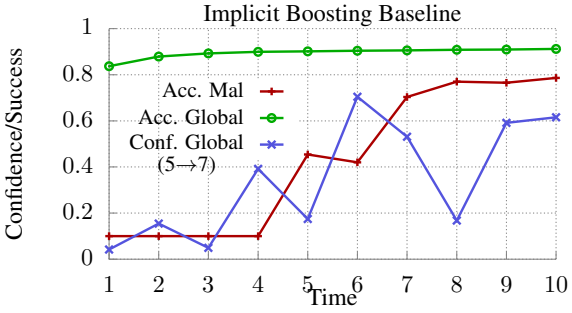


Figure 1. Implicit boosting attack metrics

2. Implicit Boosting

While the loss is a function of a weight vector $\mathbf{w}$, we can use the chain rule to obtain the gradient of the loss with respect to the weight update $\boldsymbol{\delta}$, i.e. $\nabla_{\boldsymbol{\delta}} L = \alpha_m \nabla_{\mathbf{w}} L$. Then, initializing $\boldsymbol{\delta}$ to some appropriate $\boldsymbol{\delta}_{\text{ini}}$, the malicious agent can directly minimize with respect to $\boldsymbol{\delta}$. However, the baseline attack using *implicit boosting* (Figure 1) is much less successful than the explicit boosting baseline, with the adversarial objective only being achieved in 4 of 10 iterations. Further, it is computationally more expensive, taking an average of 2000 steps to converge at each time step, which is about $4\times$ longer than a benign agent. Since consistently delayed updates from the malicious agent might lead to it being dropped from the system in practice, we focused on explicit boosting attacks throughout.

3. Further results

Our technical report (Bhagoji et al., 2018) contains extended results beyond those presented here. These are for the CIFAR-10 dataset as well as with replicates for each time step and with larger auxiliary datasets for the Fashion-MNIST data.

3.1. Results on Adult Census dataset

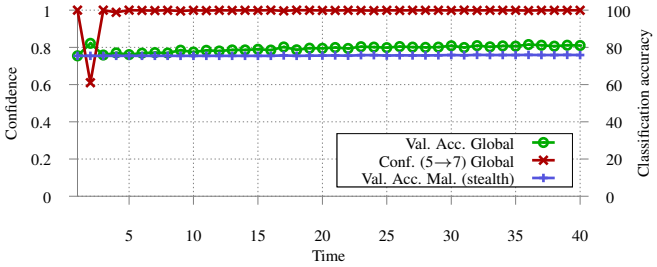
Results for the 3 different attack strategies on the Adult Census dataset (Figure 2) confirm the broad conclusions we derived from the Fashion MNIST data. The baseline attack is able to induce high confidence targeted misclassification for a random test example but affects performance on the benign objective, which drops from 84.8% in the benign case to just around 80%. The alternating minimization attack is able to ensure misclassification with a confidence of around 0.7 while maintaining 84% accuracy on the benign objective.

3.2. Multiple instance poisoning

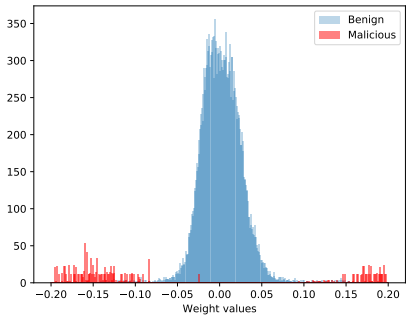
For completeness, we provide results for the case with $r = 10$, i.e. the case when the malicious agent wishes to classify 10 different examples in specific target classes. These results are given Figures 3a (targeted model poisoning) and 3b (Alternating minimization with stealth). While targeted model poisoning is able to induce targeted misclassification, it has an adverse impact on the global model's accuracy. This is countered by the alternating minimization attack, which ensures that the global model converges while still meeting the malicious objective.

[*]Work done at I.B.M. Research [1]Princeton University [2]I.B.M. T.J. Watson Research Center. Correspondence to: Arjun Nitin Bhagoji <abhagoji@princeton.edu>.

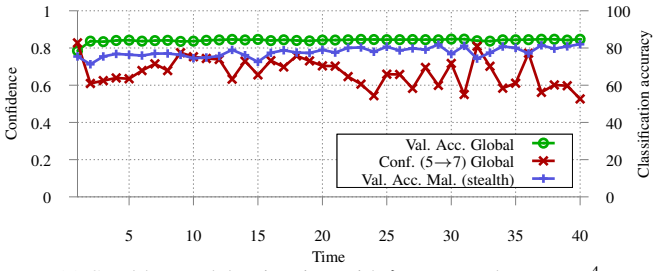
Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).



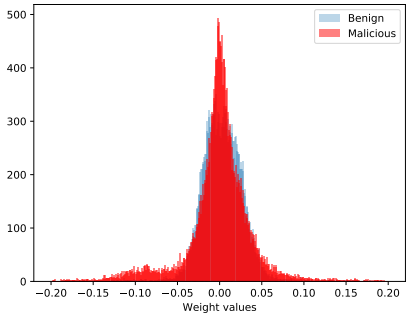
(a) Targeted model poisoning



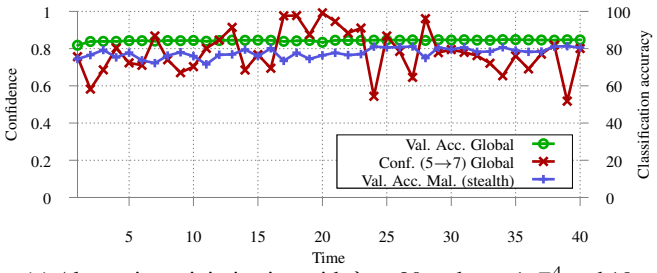
(b) Comparison of weight update distributions for targeted model poisoning



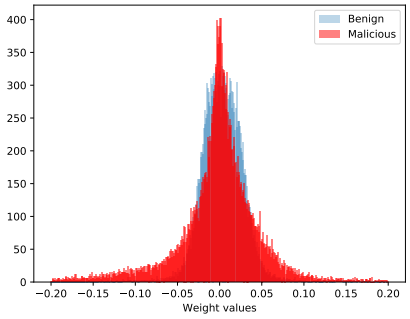
(c) Stealthy model poisoning with $\lambda = 20$ and $\rho = 1e^{-4}$



(d) Comparison of weight update distributions for stealthy model poisoning

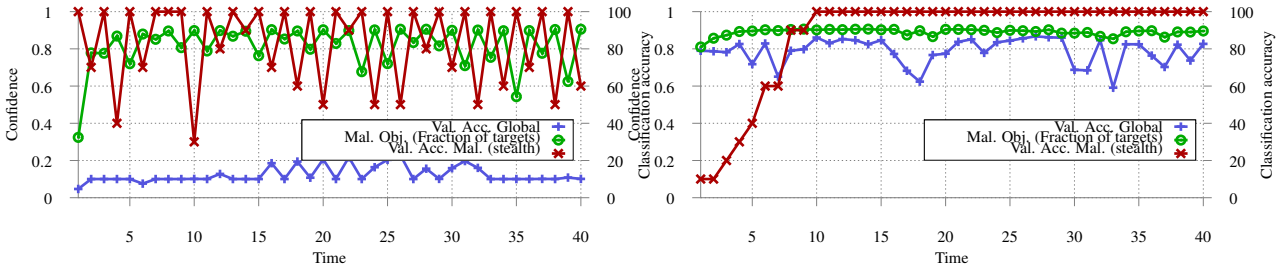


(e) Alternating minimization with $\lambda = 20$ and $\rho = 1e^{-4}$ and 10 epochs for the malicious agent



(f) Comparison of weight update distributions for alternating minimization

Figure 2. **Attacks on a fully connected neural network on the Census dataset.**



(a) Targeted model poisoning.

(b) Alternating minimization with 10 epochs for the malicious agent and 10 steps for the stealth objective for every step of the benign objective.

Figure 3. **Attacks with multiple targets ($r = 10$) for a CNN on the Fashion MNIST data.**

3.3. Randomized agent selection

When the number of agents increases to $k = 100$, the malicious agent is not selected in every step. Further, the size of $|\mathcal{D}_m|$ decreases, which makes the benign training step in the alternating minimization attack more challenging. The challenges posed in this setting are reflected in Figure 4a, where although targeted model poisoning is able to introduce a targeted backdoor, it is not present for every step as there are steps where only benign agents provide updates. Nevertheless, targeted model poisoning is effective overall, with the malicious objective achieved along with convergence of the global model at the end of training. The alternating minimization attack strategy with stealth (Figure 4b) is also able to introduce the backdoor, as well as increase the classification accuracy of the malicious model on test data. However, the improvement in performance is limited by the paucity of data for the malicious agent. It is an open question if data augmentation could help improve this accuracy.

3.4. Bypassing Byzantine-resilient aggregation mechanisms

In Section 4, we presented the results of successful attacks on two different Byzantine resilient aggregation mechanisms: Krum (Blanchard et al., 2017) and coordinate-wise median (coomed) (Yin et al., 2018). In this section, we present the results for targeted model poisoning when Krum is used (Figure 5a). The attack uses a boosting factor of $\lambda = 2$ with $k = 10$. Since there is no need to overcome the constant scaling factor α_m, the attacks can use a much smaller boosting factor λ to ensure the global model has the targeted backdoor. With the targeted model poisoning attack, the malicious agent's update is the one chosen by Krum for 34 of 40 time steps but this causes the validation accuracy on the global model to be extremely low. Thus, our attack causes Krum to converge to an ineffective model, in contrast to its stated claims of being Byzantine-resilient. However, our attack does not achieve its goal of ensuring that the global model converges to a point with good performance on the test set due to Krum selecting just a single agent at each time step.

We also consider the effectiveness of the alternating minimization attack strategy when coomed is used for aggregation. While we have shown targeted model poisoning to be effective even when coomed is used, Figure 5b demonstrates that alternating minimization, which ensures that the local model learned at the malicious agent also has high validation accuracy, is not effective.

4. Visualization of weight update distributions

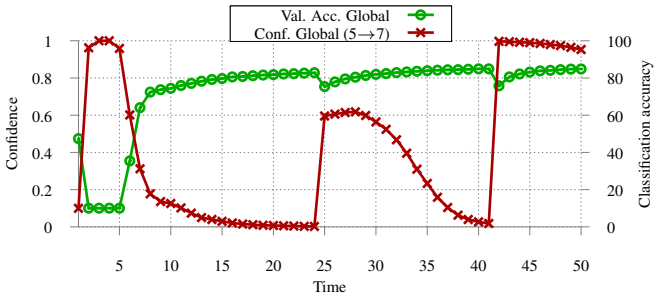
Figure 4 shows the evolution of weight update distributions for the 4 different attack strategies on the CNN trained on the Faishon MNIST dataset. Time slices of this evolution were shown in the main text of the paper. The baseline and concatenated training attacks lead to weight update distributions that differ widely for benign and malicious agents. The alternating minimization attack without distance constraints reduces this qualitative difference somewhat but the closest weight update distributions are obtained with the alternating minimization attack with distance constraints.

5. Interpretability for benign inputs

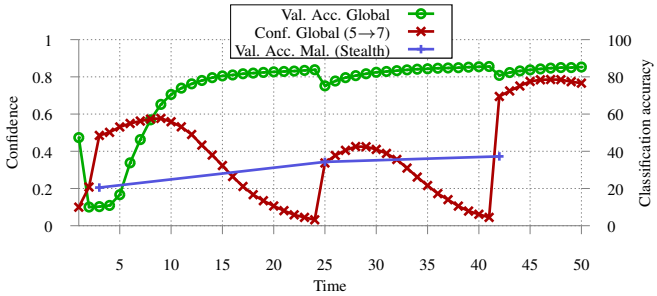
We provide additional interpretability results for global models trained with and without the presence of a malicious agent on benign data in Figures 7 and 8 respectively. These show that the presence of the malicious agent using targeted model poisoning does not significantly affect how the global model makes decisions on benign data.

References

Bhagoji, A. N., Chakraborty, S., Mittal, P., and Calo, S. Analyzing federated learning through an adversarial lens.

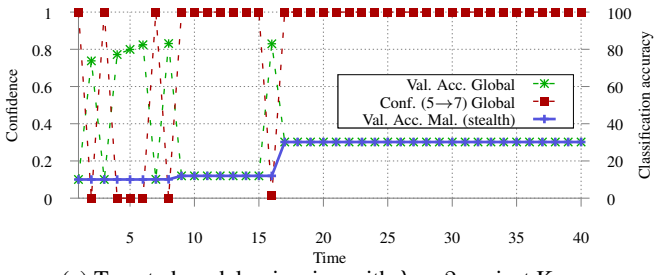


(a) Targeted model poisoning with $\lambda = 100$.

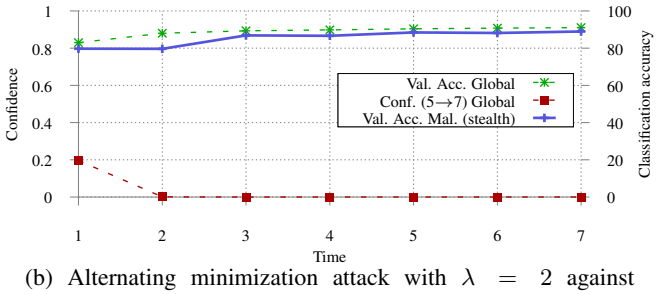


(b) Alternating minimization with $\lambda = 100$, 100 epochs for the malicious agent and 10 steps for the stealth objective for every step of the benign objective.

Figure 4. **Attacks on federated learning in a setting with $K = 100$ and a single malicious agent for a CNN on the Fashion MNIST data.**



(a) Targeted model poisoning with $\lambda = 2$ against Krum.



(b) Alternating minimization attack with $\lambda = 2$ against coomed.

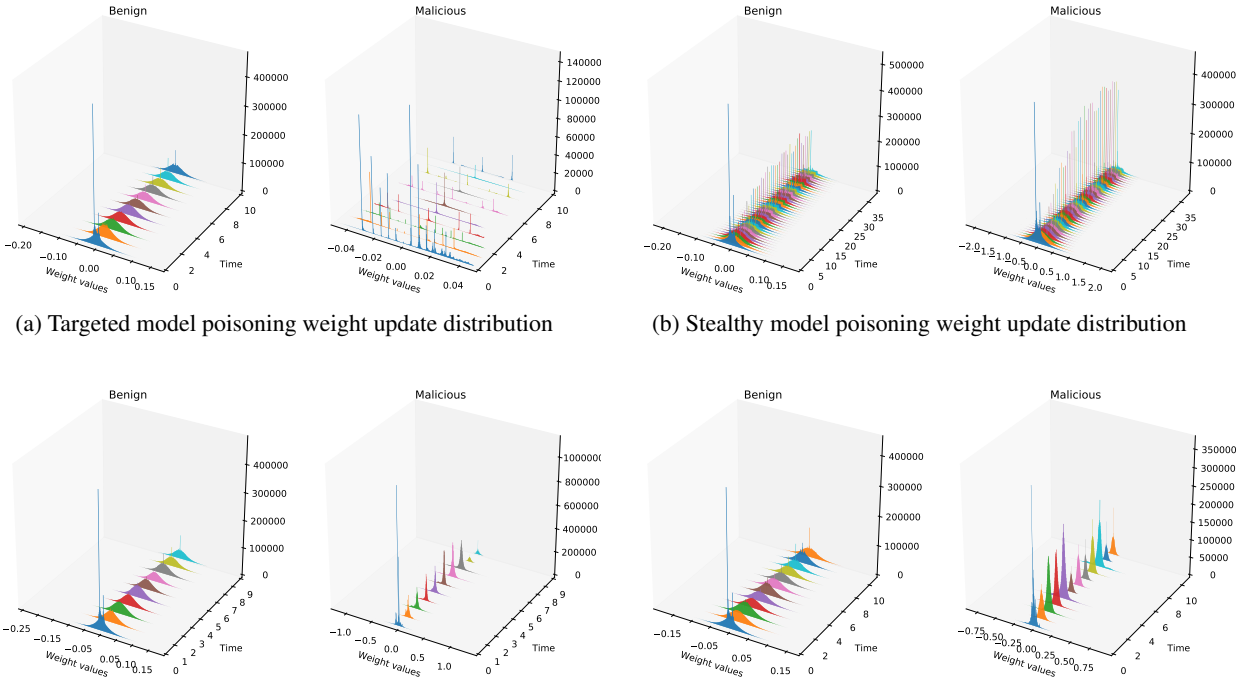
Figure 5. **Additional results for attacks on Byzantine-resilient aggregation mechanisms.**

arXiv preprint arXiv:1811.12470, 2018.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 2017.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yin, D., Chen, Y., Ramchandran, K., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498*, 2018.



(a) Targeted model poisoning weight update distribution

(b) Stealthy model poisoning weight update distribution

(c) Alternating minimization (only loss stealth) weight update distribution

(d) Alternating minimization (both loss and distance stealth) distance constraints weight update distribution

Figure 6. Weight update distribution evolution over time for all attacks on a CNN for the Fashion MNIST dataset.

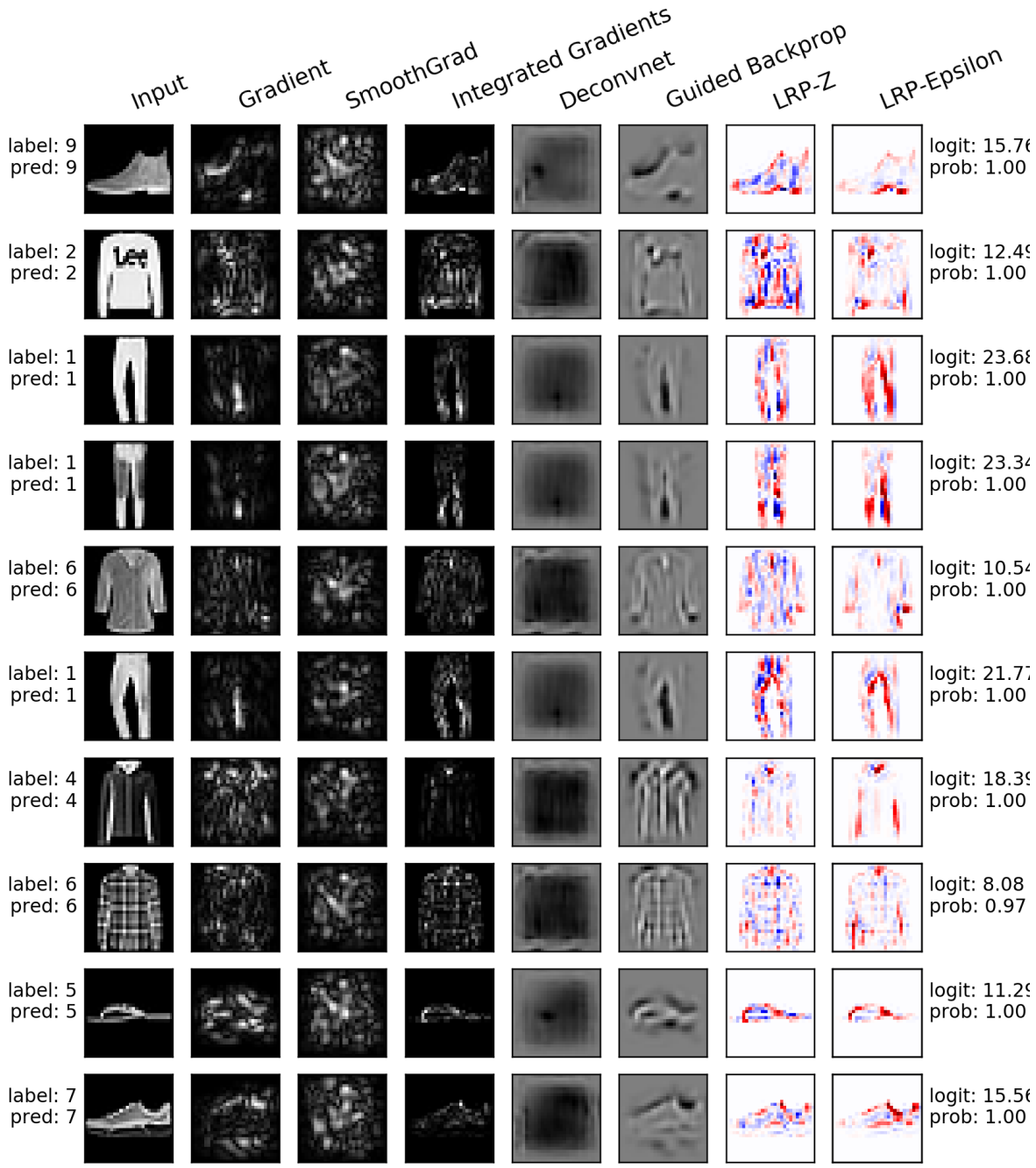


Figure 7. Decision visualizations for global model trained on Fashion MNIST data using only benign agents *on benign data.*

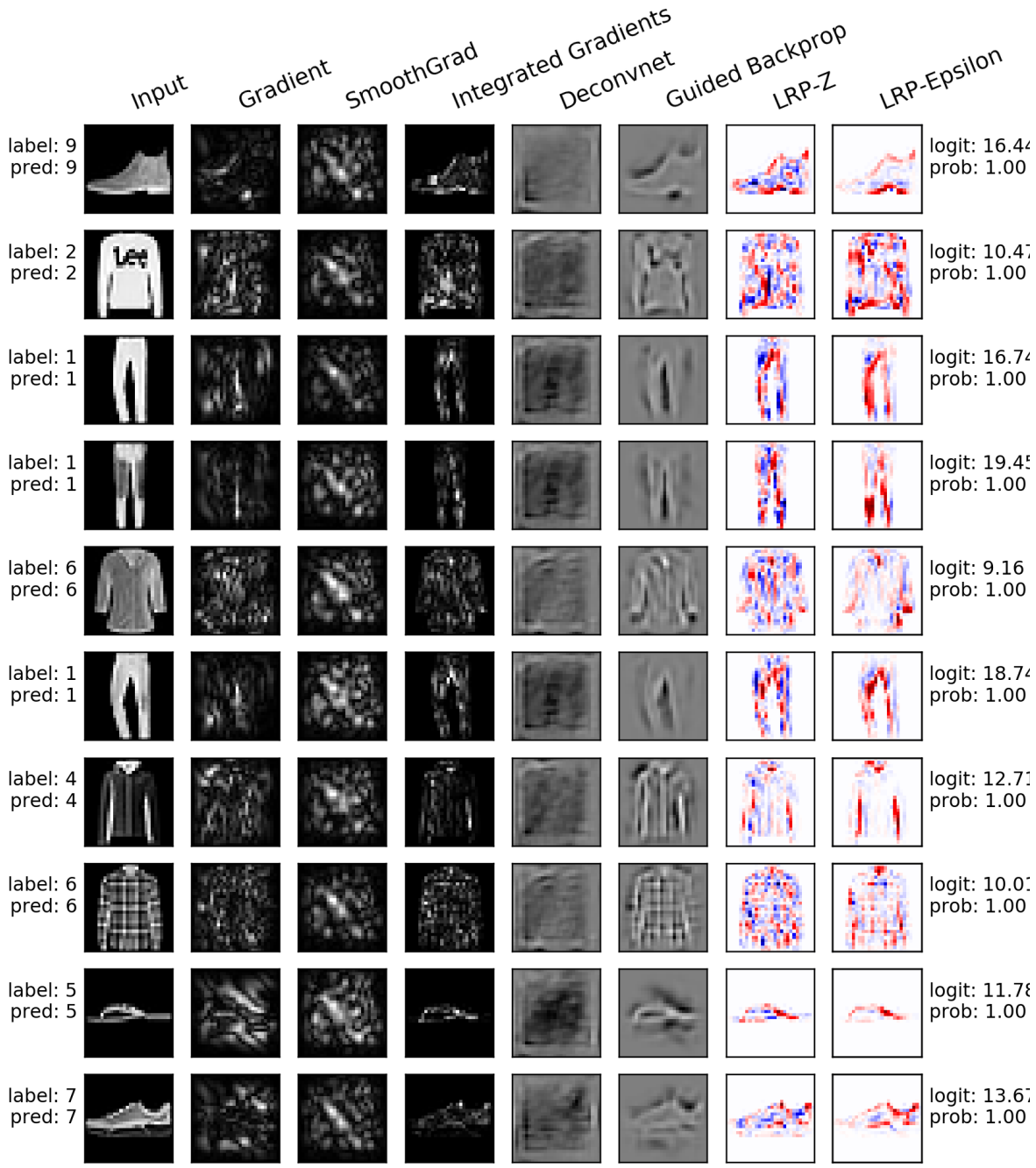


Figure 8. Decision visualizations for global model trained on Fashion MNIST data using 9 benign agents and 1 malicious agent using the baseline attack *on benign data.*