Appendix A: Targeted samples with attackers

For a better understanding of the attackers' data, we present Figure 1 to depict some examples (images) which were targeted by the label-flipping or backdoor attackers.

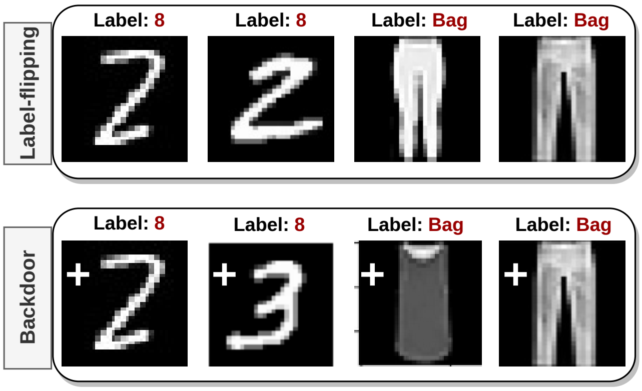


Figure 1: Some affected samples of attackers' local dataset.

Appendix B: Choosing appropriate λ

We carry out experiments with $\lambda \in \{0.5, 1, 1.5, \cdots, 4.5, 5\}$, and results are reported in Figure 2 for MNIST and Fashion-MNIST dataset. With $\lambda = 3$ for MNIST dataset, the global model provides maximum accuracy and minimum variance across all the benign clients. Similarly, $\lambda = 4.5$ can be used for Fashion-MNIST dataset. Note that the value of λ depends on the data.

Appendix C: Special cases of non-IID data

In order to validate the strength of our fairness scheme, we carry out some experiments for special cases of non-IID data without attackers. Each client receives a varying number of samples (images) of a particular number of classes. Table 1 presents the obtained results which clearly indicate that our scheme loses marginally on the accuracy and variance. Even for a diverse dataset like Fashion-MNIST, the variance is 42.5 with an extreme case (i.e., 2 classes/client) of non-IID, which is making our scheme suitable for any FL application that deems to have a fair model for new users.

Table 1: Results for special cases of non-IID data across clients in no-attacker scenario.

# Classes/ client	MNIST dataset		Fashion-MNIST dataset	
	Acc. (%)	Var	Acc (%)	Var
2	96.2	32.5	86.4	42.5
4	96.8	24.2	85.6	32.2
6	95.9	22.3	88.5	27.4
8	97.4	21.3	87.3	19.4
10	97.0	21.1	89.4	18.3

Appendix D: Combining existing defenses with existing fairness schemes

Linear combination of existing algorithms would include a defense method followed by a fairness handling method where the former needs to execute completely (all rounds) before the latter one. Table 2 reports the experimental results for the case with 100 clients (60 benign and 40 attackers). The results do not favor the linear combination as the obtained ASR and variance are higher than FFL+AD.

Table 2: Results for the linear combination of existing algorithms. [D1: FoolsGold, D2: FedAvg-RLR, F1: q-FFL, F2: AFL]

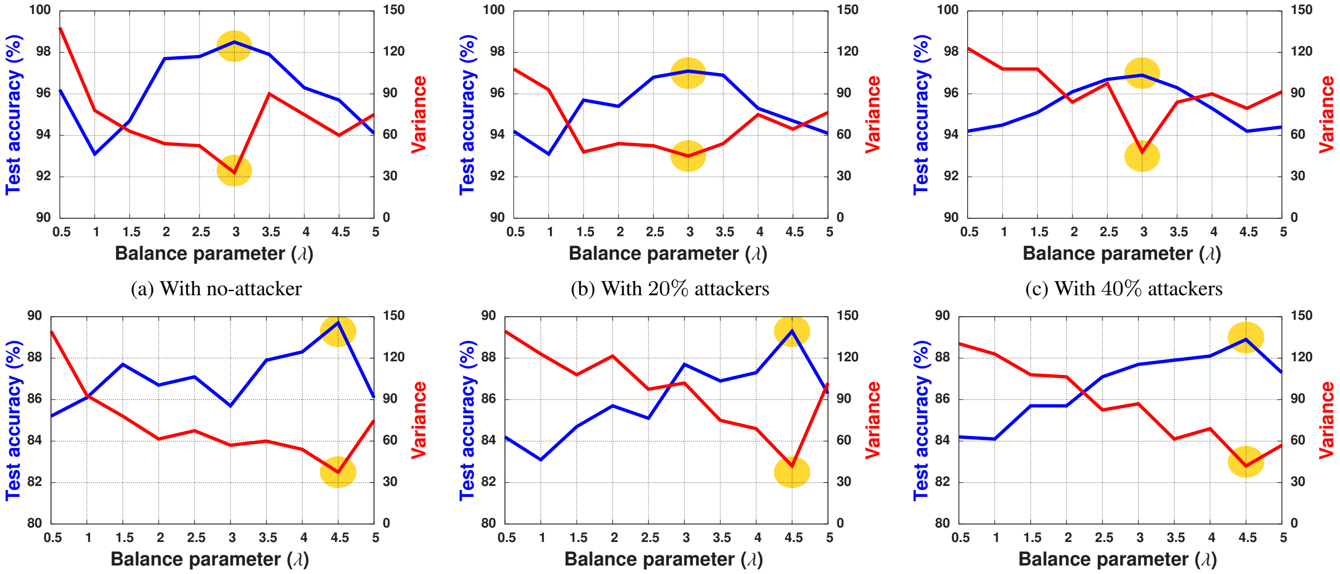
Combination	MNIST dataset		Fashion-MNIST dataset	
	ASR	Var	ASR	Var
D1+F1	0.20	115.3	0.13	122.5
D1+F2	0.20	102.1	0.13	127.3
D2+F1	0.24	106.4	0.29	114.2
D2+F2	0.24	120.3	0.29	124.5
FFL+AD	0.04	28.7	0.03	23.5

Appendix E: Training (execution) time analysis

Finally, we analyze the FFL+AD against the considered approaches using execution time. Let E_p and E_e respectively denote the training time taken by FFL+AD and an existing approach for the case with 100 clients of which 40 are attackers. For better understanding, we compute a normalized time difference $\frac{E_e - E_p}{\max(E_e, E_p)}$ that shows how fast (positive value) or slow (negative value) the FFL+AD is from the existing algorithm, and results are reported in Table 3.

Table 3: Demonstrating the normalized difference of training (execution) time taken by existing algorithms against FFL+AD.

Existing scheme	MNIST dataset	Fashion-MNIST dataset
FoolsGold	-0.19	-0.12
FedAvg-RLR	-0.09	-0.07
q-FFL	+0.28	+0.31
AFL	+0.13	+0.11



(a) With no-attacker

(b) With 20% attackers

(c) With 40% attackers

(d) With no-attacker

(e) With 20% attackers

(f) With 40% attackers

Figure 2: Accuracy and variance with varying λ for MNIST dataset in (a), (b), and (c), and for Fashion-MNIST in (d), (e), and (f). At $\lambda = 4.5$, accuracy and variance are optimum. Read the x-axis with *left* y-axis (blue line) for accuracy and with *right* y-axis for variance (red line).