A Federated learning pipeline

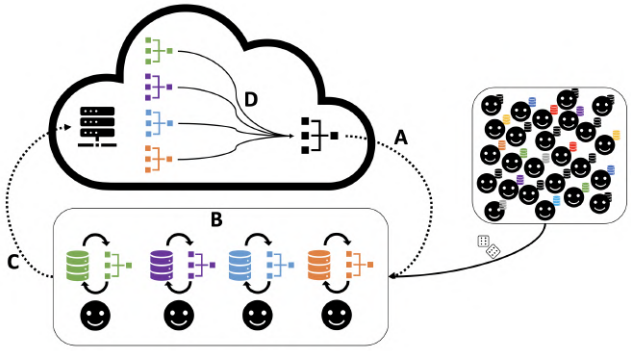

Figure 5: Federated Learning pipeline. **A:** The global model is sent to a random subset of clients. **B:** The clients train the model locally using their data. **C:** The local models are sent to an aggregation server. **D:** The client models are aggregated to make a new global model. This process is repeated until convergence.

B Robust aggregation methods

Name	Description	Robust	Server data
FA (or FedAvg)	Federated Averaging [6]	No	No
COMED	Coordinate-wise median [7]	Yes	No
MKRUM	Multi-Krum [9]	Yes	No
AFA	Adaptive Federated Averaging [10]	Yes	No
FedMGDA++	FedMGDA+ [11]	Yes	No
FedDF	FedDF [16]	No	Yes
FedDFmed	FedDF with median-based Knowledge Distillation	Yes	Yes
FedBE	FedBE [17]	No	Yes
FedBEmed	FedBE with median-based Knowledge Distillation	Yes	Yes
FedADF	AFA [10] with an additional median-based Knowledge Distillation	Yes	Yes
FedMGDA+DF	FedMGDA+ [11] with an additional median-based Knowledge Distillation	Yes	Yes
FedRAD	Our novel aggregator, see Algorithm 2	Yes	Yes
FedRADnoise	FedRAD using uniform noise instead of server-side data	Yes	No

Table 2: List of aggregation methods. Methods in **bold** are novel. "Robust" is used to describe whether steps have been taken to detect or defend against adversaries. "Server data" is used to indicate whether an *unlabelled* dataset is required on the server side.

C Robustness of the median

Table 3 illustrates the advantage of using a robust statistic such as the median instead of mean for model aggregation.

	Logits predictions from 10 clients	Average	Median
No attacker	$[1, 1, 2, 2, 3, 3, 4, 4, 5, 5]$	3	3
One weak attacker	$[1, 1, 2, 2, 3, 3, 4, 4, 5, 15]$	4	3
Four weak attackers	$[1, 1, 2, 2, 3, 3, 14, 14, 15, 15]$	7	3
One strong attackers	$[1, 1, 2, 2, 3, 3, 4, 4, 5, 1005]$	103	3
Four strong attackers	$[1, 1, 2, 2, 3, 3, 1004, 1004, 1005, 1005]$	403	3

Table 3: An illustrative example demonstrating the robustness of the median against attackers with confidently incorrect predictions. The logits values are hypothetical outputs from 10 different models for a particular class. Attackers are coloured in red.

D Learning curves of attacks

D.1 No attacks

Results from experiments with no attacks for different levels of data heterogeneity can be seen in Figure 6.

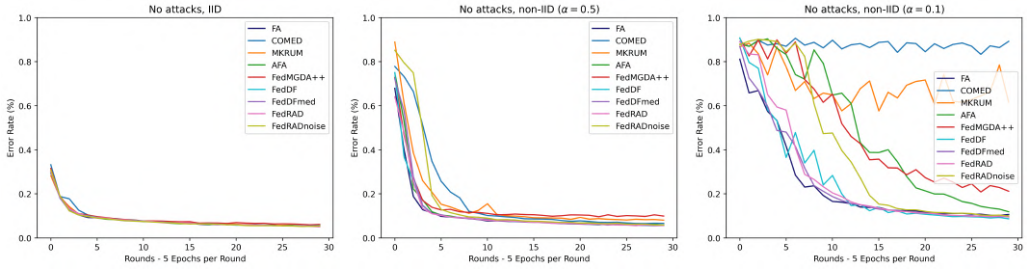

Figure 6: Effects of non-IID data for no attacks.

All the methods perform roughly the same on IID data when there are no attackers. With increased heterogeneity, the learning starts to slow down for all aggregators. When data is very non-IID, i.e. $\alpha = 0.1$, the performance of aggregators such as COMED, MKRUM, AFA and FedMGDA+ declines significantly. Both FedMGDA+ and AFA slow down quite a bit, and both of them block some healthy clients which effectively decreases the size of the training set.

Federated Averaging still performs well in non-IID circumstances when there are no attackers. Methods that use Knowledge Distillation also perform similarly to FedAvg.

D.2 Faulty Attacks

Faulty attackers, also known as Byzantine, are attackers which update their models by adding a lot of random noise to the model parameters. Results for 1, 5 and 10 faulty attackers can be seen in Figures 7, 8 and 9 respectively.

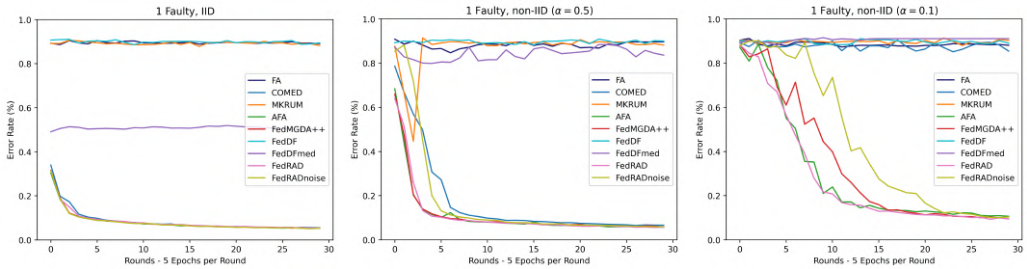

Figure 7: Effects of non-IID data with 1 Faulty attacker.

Just 1 faulty attacker completely ruins performance for non-robust aggregators. Both AFA and FedMGDA+ are able to effectively block the attacker and learn, even with non-IID data.

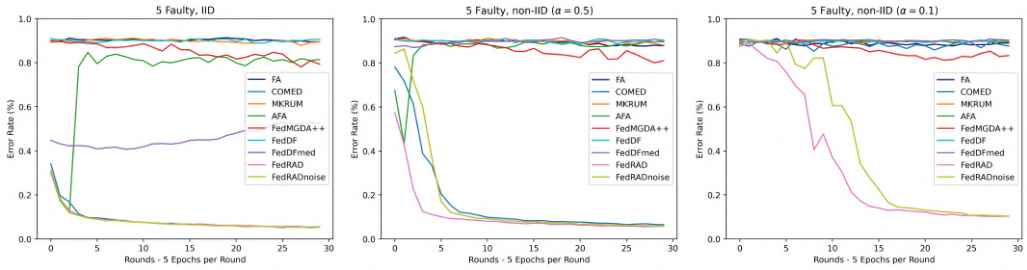

Figure 8: Effects of non-IID data with 5 Faulty attackers.

Once more attackers are added, AFA and FedMGDA+ start to fail. COMED still works for homogeneous data. The only aggregators that work against 5 faulty clients in the most heterogeneous case are our novel methods.

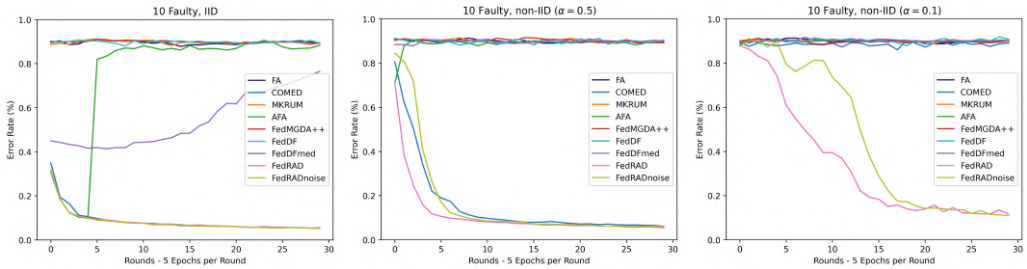

Figure 9: Effects of non-IID data with 10 Faulty attackers.

With 10 attackers and non-IID data FedRAD is the only aggregator that manages to learn anything. Even FedRADnoise, which uses only random noise for scoring and distillation, is able to learn well. The good performance of aggregators which use our novel median-scoring mechanism against is explained by Figure 1, which shows that Faulty models rarely ever give the median logit prediction for a given class, and thus often get a score of 0, which is equivalent to leaving faulty models out completely during the weighted averaging.

D.3 Malicious Attacks

Malicious attackers are agents who train the models with incorrect labels. In our case, all labels are set to 0 before training. Results for 1, 5, and 10 attackers can be seen in Figures 10, 11 and 12 respectively.

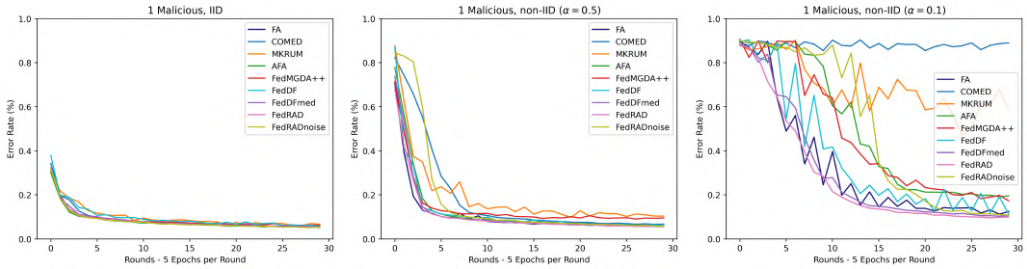

Figure 10: Effects of non-IID data with 1 malicious attacker.

One malicious attacker does not impact learning very much.

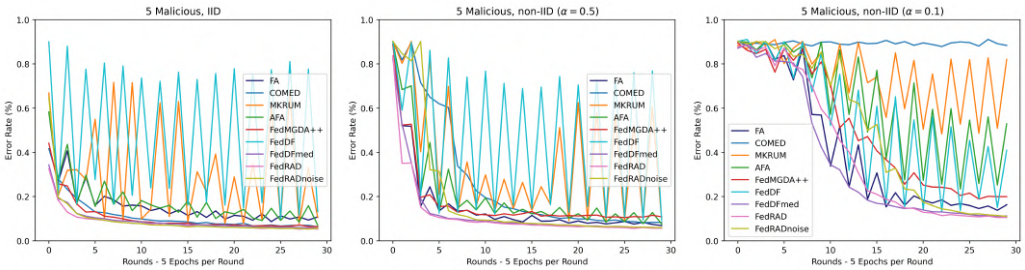

Figure 11: Effects of non-IID data with 5 malicious attackers.

With 5 malicious attackers, many aggregators start to perform badly and the training becomes very noisy. This is where the benefits of median-based Knowledge distillation starts to shine through: FedDF training is very noisy, but FedDFmed is among the best performing methods.

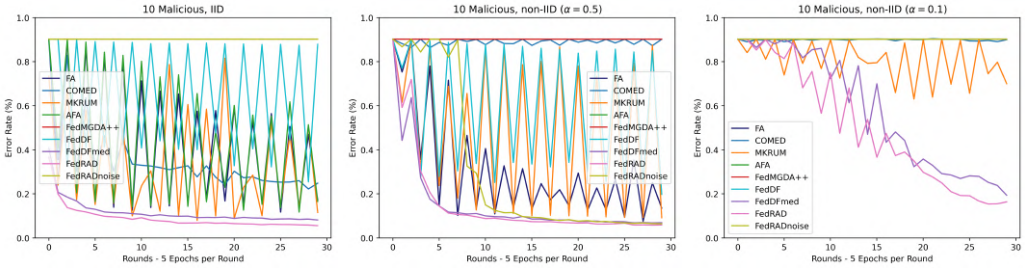

Figure 12: Effects of non-IID data with 10 malicious attackers.

With 10 malicious attackers, even more aggregators start to fail. Robust methods which rely on comparing model parameters, such as AFA and FedMGDA+, don't learn anything at all in the non-IID case.

The best performing aggregators against malicious attacks in both IID and non-IID scenarios are our novel methods which utilize median-based Knowledge Distillation: FedDFmed and FedRAD. This is explained by the median-counting histograms in Figure 1, which show that median logits are less contaminated by malicious agents.

MKRUM, AFA and FedMGDA+ fail because models from malicious agents are quite similar to healthy models when comparing their high-dimensional parameters. In non-IID situations, it becomes difficult to detect malicious models among the non-IID healthy models by using distance metrics on their parameters.

D.4 Both Types of Attacks

Experiments are done using both faulty and malicious attackers in IID and non-IID settings. The results from these can be seen in Figures 4, and 13.

Using 10 faulty, 10 malicious and only 10 honest agents, we finally reached the breaking point. But even with this many attakcers, our FedRAD aggregator was the only one that showed any progress and managed to reach a 60% error rate on IID data, which is respectable considering only a third of the agents are honest.

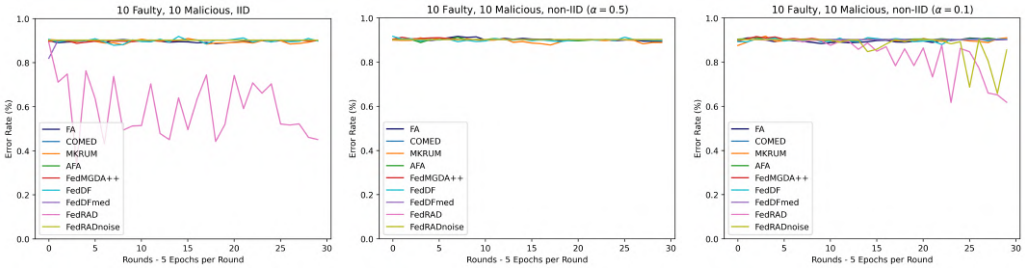

Figure 13: Effects of non-IID data with 10 faulty and 10 malicious attackers.

E Median-based knowledge distillation for different aggregators

The median based knowledge distillation module is also tested with AFA and FedMGDA, called FedADF and FedMGDA+DF. FedBE and FedBEmed (which uses median-logits) were also tested. Results are shown in Table 4.

Attacks	Aggregator	Error rates (%)		
		IID	non-IID	
			$\alpha = 0.5$	$\alpha = 0.1$
No attacks	AFA	$\mathbf{5.28 \pm 0.17}$	5.94 ± 0.16	$\mathbf{10.36 \pm 0.96}$
	FedMGDA++	6.28 ± 0.25	12.15 ± 5.39	21.12 ± 3.84
	FedBE	$\mathbf{5.36 \pm 0.13}$	$\mathbf{5.93 \pm 0.22}$	11.70 ± 1.38
	FedBEmed	5.39 ± 0.10	$\mathbf{5.85 \pm 0.28}$	$\mathbf{9.67 \pm 0.29}$
	FedADF	5.45 ± 0.14	6.38 ± 0.31	16.81 ± 4.55
	FedMGDA+DF	6.27 ± 0.32	8.29 ± 0.45	15.11 ± 0.57
	FedRAD	$\mathbf{5.29 \pm 0.16}$	$\mathbf{5.82 \pm 0.15}$	$\mathbf{9.38 \pm 0.57}$
	FedRADnoise	$\mathbf{5.32 \pm 0.15}$	5.93 ± 0.09	$\mathbf{9.71 \pm 0.57}$
10 Faulty	AFA	70.97 ± 32.56	90.12 ± 0.47	89.70 ± 0.33
	FedMGDA++	89.82 ± 0.71	89.37 ± 0.28	89.89 ± 0.43
	FedBE	90.24 ± 0.85	90.02 ± 1.14	89.87 ± 0.55
	FedBEmed	90.30 ± 0.86	90.27 ± 0.72	89.40 ± 0.68
	FedADF	$\mathbf{5.44 \pm 0.18}$	6.45 ± 0.36	27.68 ± 30.61
	FedMGDA+DF	54.54 ± 8.31	90.11 ± 0.32	90.19 ± 0.46
	FedRAD	$\mathbf{5.44 \pm 0.07}$	$\mathbf{5.91 \pm 0.31}$	$\mathbf{12.63 \pm 3.65}$
	FedRADnoise	$\mathbf{5.29 \pm 0.21}$	$\mathbf{5.99 \pm 0.08}$	$\mathbf{13.07 \pm 3.29}$
10 Malicious	AFA	17.68 ± 4.46	90.20 ± 0.00	90.20 ± 0.00
	FedMGDA++	90.20 ± 0.00	90.20 ± 0.00	90.20 ± 0.00
	FedBE	30.40 ± 3.75	33.50 ± 9.16	29.53 ± 3.10
	FedBEmed	25.14 ± 4.03	63.77 ± 32.99	54.21 ± 34.26
	FedADF	13.11 ± 2.67	90.20 ± 0.00	84.83 ± 10.74
	FedMGDA+DF	70.55 ± 7.44	67.68 ± 11.76	86.42 ± 5.78
	FedRAD	$\mathbf{5.89 \pm 0.20}$	$\mathbf{6.55 \pm 0.43}$	$\mathbf{13.09 \pm 1.97}$
	FedRADnoise	90.20 ± 0.00	73.45 ± 33.50	59.49 ± 37.61
5 Faulty, 5 Malicious	AFA	67.62 ± 6.64	67.82 ± 30.59	80.97 ± 17.03
	FedMGDA++	90.20 ± 0.00	90.21 ± 0.00	89.31 ± 1.61
	FedBE	90.36 ± 0.28	90.47 ± 0.51	90.25 ± 0.32
	FedBEmed	90.21 ± 1.21	90.25 ± 0.42	89.84 ± 0.31
	FedADF	6.01 ± 0.19	9.53 ± 3.76	19.47 ± 8.33
	FedMGDA+DF	47.88 ± 4.25	87.82 ± 2.87	89.91 ± 0.82
	FedRAD	$\mathbf{5.63 \pm 0.20}$	$\mathbf{6.15 \pm 0.40}$	$\mathbf{10.95 \pm 1.22}$
	FedRADnoise	$\mathbf{5.64 \pm 0.17}$	23.39 ± 33.41	12.84 ± 1.52

Table 4: Test set error rate for MNIST after 30 rounds. Average and standard deviation of test-set error rates shown for 5 different random seeds.