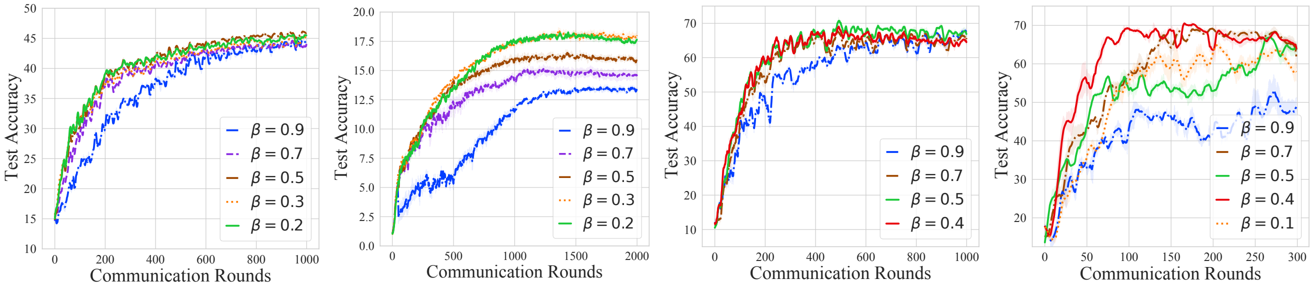
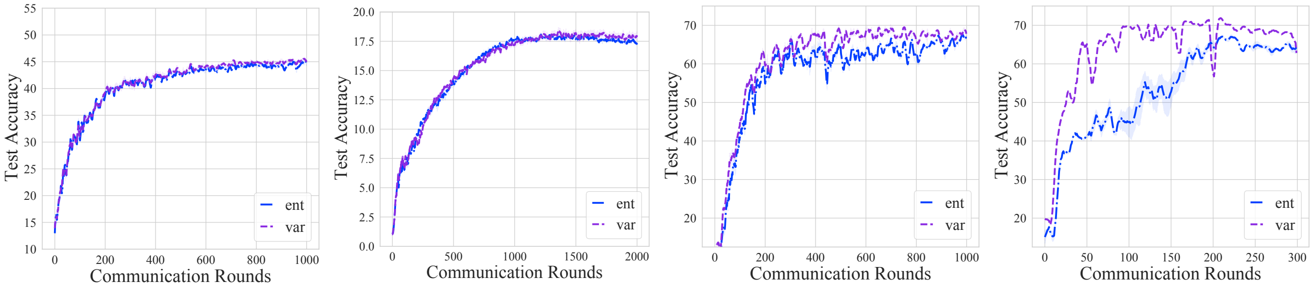


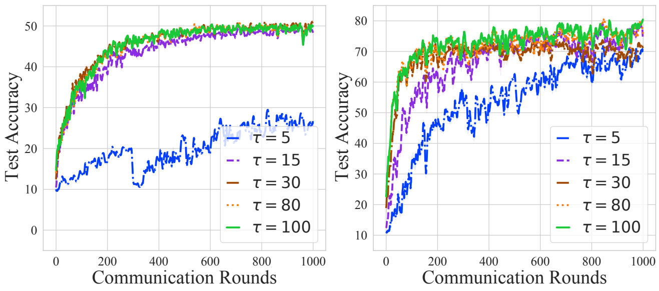
(a) CIFAR10 (b) CIFAR100 (c) OrganAMNIST (d) BloodMNIST

Figure 7: Ablation study of FEDLABEL's performance on the thresholding parameter ($0 \leq \beta \leq 1$ in (4)). For all the datasets, the performance of FEDLABEL is the lowest for the highest $\beta = 0.9$ meaning that for a too high β parameter, FEDLABEL filters out too much unlabeled data that the model is not able to learn much. FEDLABEL achieves the best performance for a lower β of the range $0.3 \sim 0.5$.



(a) CIFAR10 (b) CIFAR100 (c) OrganAMNIST (d) BloodMNIST

Figure 8: FEDLABEL's performance with different confidence metrics ($h(\cdot)$ in (4)), variance and entropy. For CIFAR10 and CIFAR100 both entropy and variance perform similarly. However, for OrganAMNIST and BloodMNIST variance performs better than entropy showing that it can better judge the confidence of the models than entropy.



(a) CIFAR10 (b) OrganAMNIST

Figure 9: Ablation study on the number of local steps τ to obtain the local model $\mathbf{w}_{\mathcal{L},k}$ for client $k \in [M]$ for larger number of labeled data (50% for CIFAR10 and 20% for OrganAMNIST). The smallest $\tau = 5$ yields the worst performance, but a slightly larger τ can largely improve the performance.

A. Additional Experimental Results

For all datasets, 80% is for training partitioned across clients and the $5\%,\ 15\%$ of the data is for the validation and test respectively. We experiment with 3 different seeds for the randomness in the dataset partition across clients and present the averaged results.

Ablation Study on Thresholding. The thresholding parameter of FEDLABEL ($0 \leq \beta \leq 1$ in (4)) determines whether either the local or global model produces a high-enough confidence logit to be used for training which has been proven to be effective for SSL [35]. We perform an ablation study on the different values of $0 \leq \beta \leq 1$ and its effect on the overall performance of FEDLABEL in Fig. 7. In Fig. 7, for all different datasets for a high $\beta = 0.9$, the test accuracy is the lowest, implying that if we set β to a too high value, FEDLABEL filters out too much unlabeled data and the model is not able to learn effectively. As we lower the threshold $\beta < 0.9$, the test accuracy improves significantly by at most approximately 17%. We observe that the best performing β value can be surprisingly low to around the range of $\beta \in [0.3, 0.5]$ for the different datasets. This is due to the RandAugumentation step which heavily transforms the image so that the image becomes different from the images the clients have been training. Hence the overall confidence on the strongly-augumented image becomes lower as observed in Fig. 7.

Gauging Confidence of the Models. One may wonder what is the appropriate metric to measure the confidence of the models' logits ($h(\cdot)$ in (3)). While we use variance as the confidence metric for all the other results in our work, we investigate how another representative measure for confidence, entropy, compares to the variance metric. Entropy is a commonly used metric to measure the uncertainty of a probability distribution, and since logits are discrete probabilities it can also be presented as an adequate candidate for measuring the confidence of the models. In Fig. 8, we show that for CIFAR10 and CIFAR100 both entropy and variance performs similarly. However, for OrganAMNIST and BloodMNIST variance performs better than entropy showing that it can better judge the confidence of the models than entropy. We conjecture that this is due to entropy compressing the confidence values to be between $[0, 1]$ while variance doesn't have this, allowing to compare the confidence of the different logits more accurately.

Number of Training Steps to Obtain the Local Model for Larger Number of Labeled Data. For a larger number of labeled data, compared to the results in Fig. 6, we see in Fig. 9 that the smallest $\tau = 5$ also yields the worst performance across the range of τ. However, even for a slightly larger τ, we can see that the performance improves with a much larger gap than the results in Fig. 6. This indicates that with a larger number of labeled data, even a slightly large τ can make the local model sufficiently represent the local data of the client, and improve the performance of FEDLABEL.