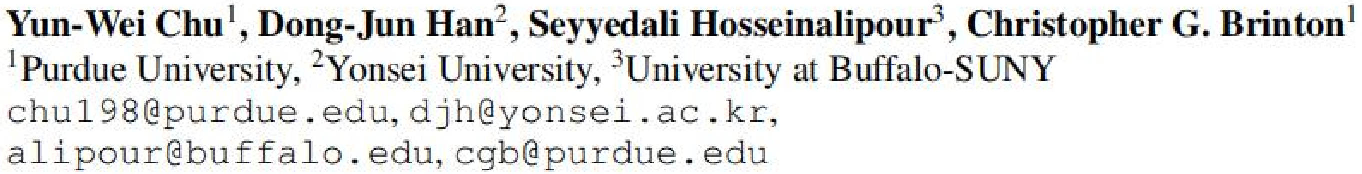
UNLOCKING THE POTENTIAL OF MODEL CALIBRATION IN FEDERATED LEARNING



Yun-Wei Chu[1], **Dong-Jun Han**[2], **Seyyedali Hosseinalipour**[3], **Christopher G. Brinton**[1]
[1]Purdue University, [2]Yonsei University, [3]University at Buffalo-SUNY
chu198@purdue.edu, djh@yonsei.ac.kr,
alipour@buffalo.edu, cgb@purdue.edu

ICLR 2025

Motivation

- Most of works in FL consider accuracy as the main performance metric.
- Beyond accuracy, in various decision-making scenarios where an incorrect prediction may result in high risk (e.g., medical applications or autonomous driving), it is also crucial for the users to determine whether to rely on the FL model’s prediction or not for each decision.
- the trained FL model should have a reliable confidence in each of its predictions, meaning that the confidence of the neural network matches well with its actual accuracy.
- In centralized settings, neural networks are often miscalibrated, indicating that the prediction confidence of the model does not accurately reflect the probability of correctness.
    - even more important in many FL use-cases
    - an overconfident global model could lead to misinformed decisions with potentially severe consequences for each client

Motivation

- In centralized training settings, research generally follows two paths to address this miscalibration issue.

  - train-time calibration methods: incorporate explicit regularizers during the training process to adjust neural networks, scaling back over/under confident predictions

  - post-hoc calibration: transforms the network's output vector to align the confidence of the predicted label with the actual likelihood of that label for the sample(applied to the already trained model to improve calibration **using an additional holdout dataset**)

  - impractical for FL

Observation

How can we unlock the potential of model calibration in FL while not sacrificing accuracy?

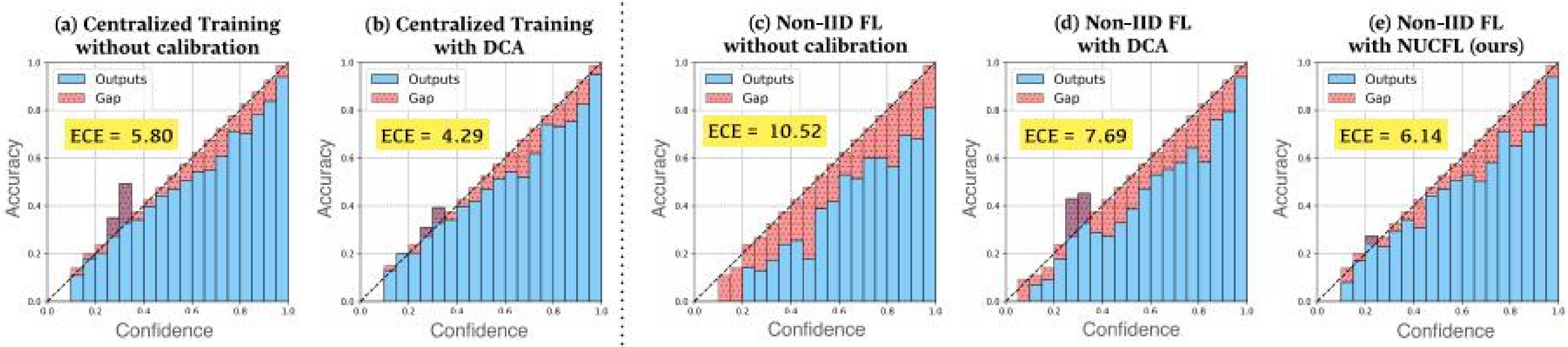
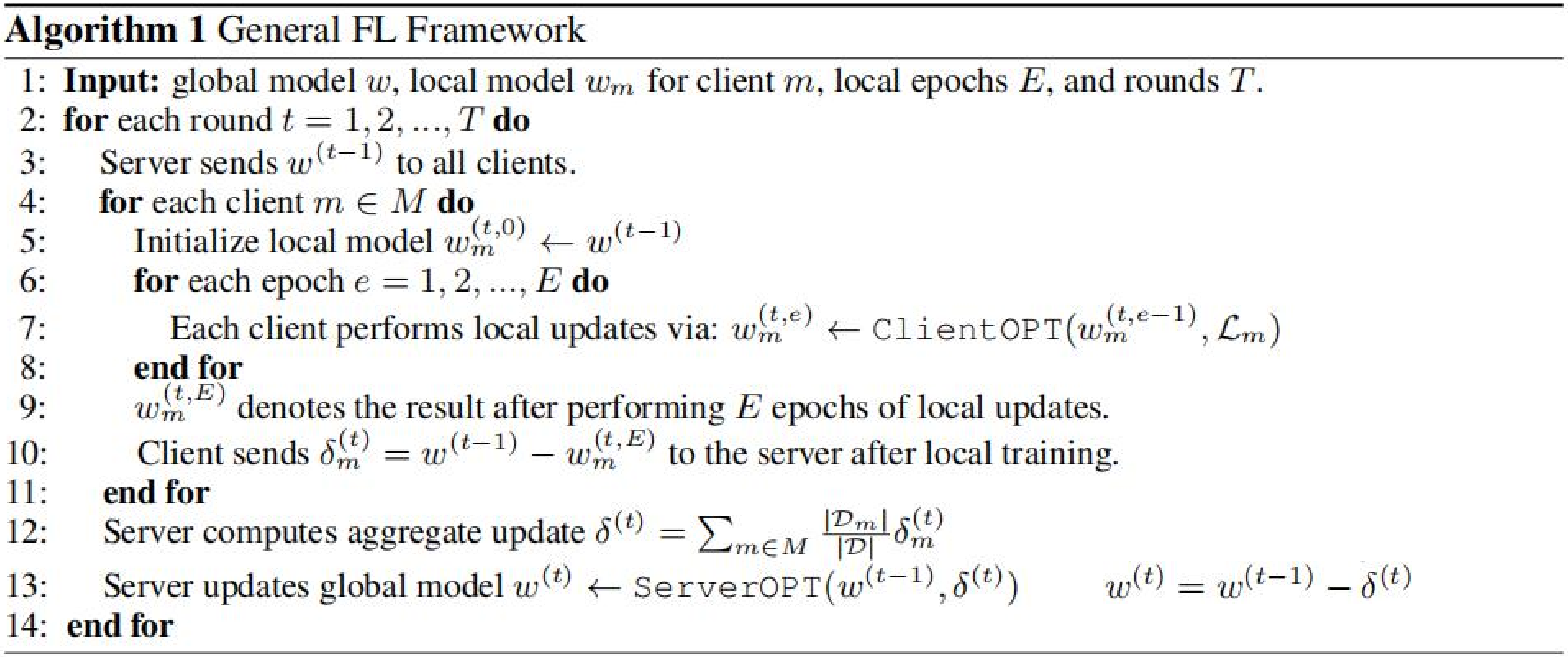


Figure 1: Reliability diagrams and calibration errors for centralized training and non-IID FL (using FedAvg) trained with various calibration methods on CIFAR-100 dataset. Our method ensures well-calibrated FL, evidenced by a notably smaller calibration error and a smaller gap (red region) between confidence and accuracy.

- FL experiences more significant model miscalibration than centralized learning.
  - This disparity is likely due to data heterogeneity across distributed FL clients, causing each client's local data to have a different impact on calibration performance.
- When applying auxiliary-based calibration methods(DCA/MDCA), a better ECE is achieved compared to FL without calibration.
  - Fig. 1(d) is still not well-calibrated by neglecting global calibration needs in heterogeneous FL settings, resulting in overconfident predictions.

Method



Algorithm 1 General FL Framework

```
Input: global model $w$, local model $w_m$ for client $m$, local epochs $E$, and rounds $T$.
for each round $t = 1, 2, ..., T$ do
    Server sends $w^{(t-1)}$ to all clients.
    for each client $m \in M$ do
        Initialize local model $w_m^{(t,0)} \leftarrow w^{(t-1)}$
        for each epoch $e = 1, 2, ..., E$ do
            Each client performs local updates via: $w_m^{(t,e)} \leftarrow \texttt{ClientOPT}(w_m^{(t,e-1)}, \mathcal{L}_m)$
        end for
        $w_m^{(t,E)}$ denotes the result after performing $E$ epochs of local updates.
        Client sends $\delta_m^{(t)} = w^{(t-1)} - w_m^{(t,E)}$ to the server after local training.
    end for
    Server computes aggregate update $\delta^{(t)} = \sum_{m \in M} \frac{|\mathcal{D}_m|}{|\mathcal{D}|} \delta_m^{(t)}$
    Server updates global model $w^{(t)} \leftarrow \texttt{ServerOPT}(w^{(t-1)}, \delta^{(t)})$        $w^{(t)} = w^{(t-1)} - \delta^{(t)}$
end for
```

Method

- BRIDGING FEDERATED LEARNING AND MODEL CALIBRATION

$$\mathcal{L}_m(w_m^{(t,e)}) = \frac{1}{|\mathcal{D}_m|}\sum_{i=1}^{|\mathcal{D}_m|} \ell(w_m^{(t,e)}; x_i, y_i),$$

- 1) integrate the calibration loss with the FL loss

$$\mathcal{L}_m^{cal}(w_m^{(t,e)}) = \frac{1}{|\mathcal{D}_m|}\sum_{i=1}^{|\mathcal{D}_m|} \left[\ell(w_m^{(t,e)}; x_i, y_i) + \beta_m \ell_{cal}(w_m^{(t,e)}; x_i, y_i)\right],$$

auxiliary calibration loss : DCA

$$\ell_{cal}\left(\{c_i\}_{1\le i\le N}, \{\hat{s}_i\}_{1\le i\le N}\right) = \ell_{dca} = \left|\frac{1}{N}\sum_{i=1}^{N} c_i - \frac{1}{N}\sum_{i=1}^{N} \hat{s}_i\right|,$$

a Multi-class DCA (MDCA) :

$$\ell_{cal}\left(\{c_i\}_{1\le i\le N}, \{s_i\}_{1\le i\le N}\right) = \ell_{mdca} = \frac{1}{K}\sum_{j=1}^{K}\left|\frac{1}{N}\sum_{i=1}^{N} c_i[j] - \frac{1}{N}\sum_{i=1}^{N} s_i[j]\right|,$$

- enhance calibration without significantly impacting the primary classification loss, ℓ.

Method

- 2) remaining challenge : Choosing an appropriate βm tailored to each client m
- Direct application of auxiliary loss calibration methods without adapting to FL characteristics would result in uniform calibration weights, such that β1 = β2 = ... = βm.
    - leads to local models being calibrated solely to their respective datasets Dm
    - uniformly large weights potentially neglecting accuracy improvements from classification
    - uniformly small weights biasing calibration toward local heterogeneity

- risks neglecting the broader global calibration needs necessary for optimal performance across the entire global distribution.

• Non-uniform penalty design(NUCFL)

- a local model closely resembling the global model is likely to represent global characteristics well, suggesting that the penalty appropriately reflects the calibration needs of the global model.
- a dissimilar/heterogeneous local model suggests a focus on local objectives (e.g., to improve accuracy) at the expense of global alignment.

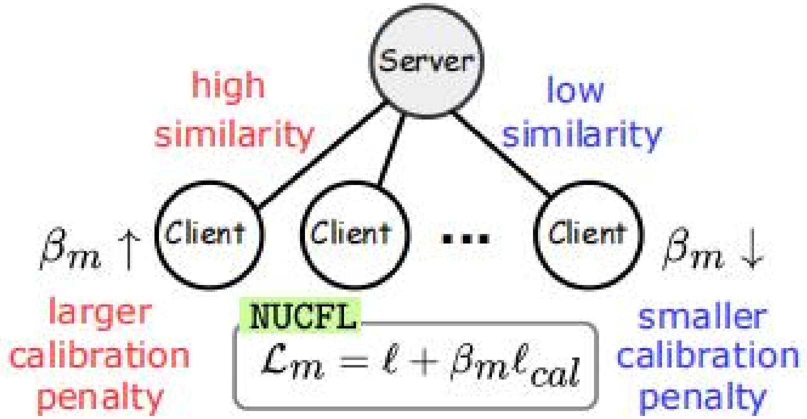
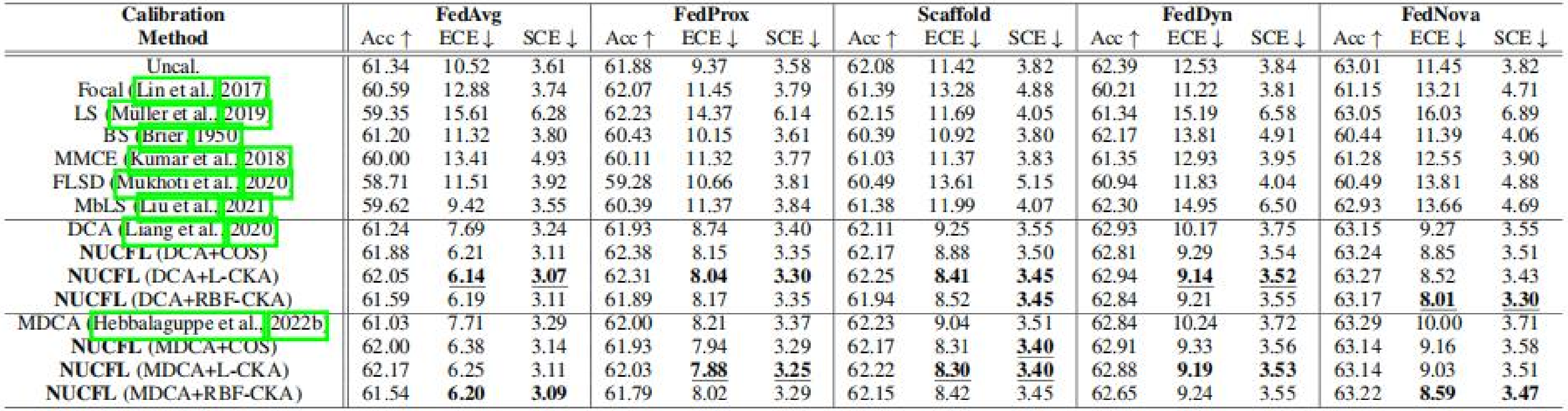


Figure 2: Idea of proposed NUCFL.

$$\beta_m = \mathtt{sim}(\delta^{(t-1)}, \delta_m^{(t,e)}), \qquad \delta_m^{(t,e)} = w^{(t-1)} - w_m^{(t,e)}$$

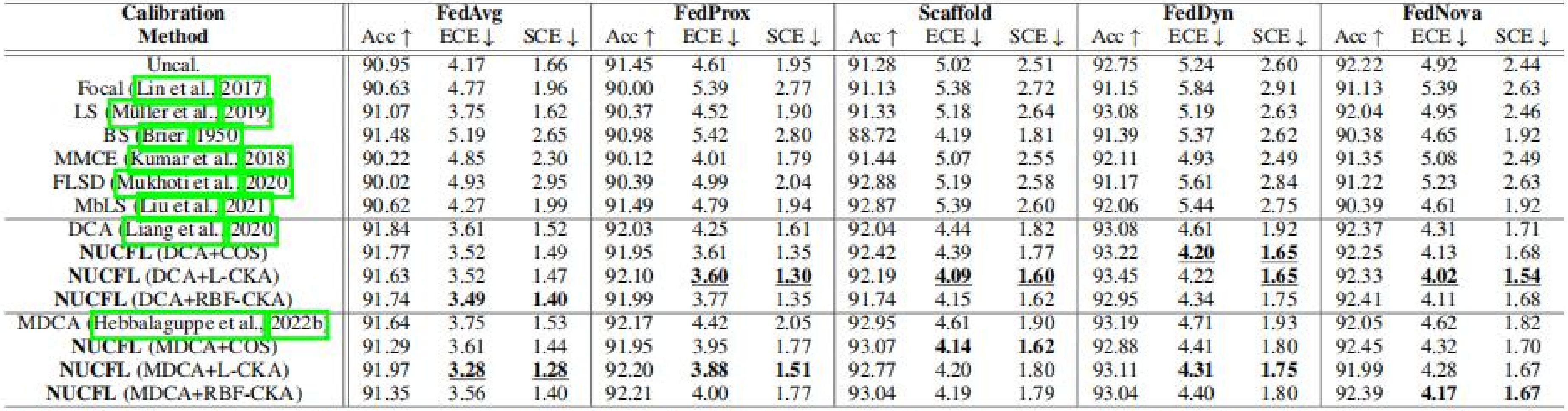
Experiments

$$\mathrm{ECE} = \sum_{i=1}^{I} \frac{B_i}{N} |A_i - C_i| \qquad \mathrm{SCE} = \frac{1}{K} \sum_{i=1}^{I} \sum_{j=1}^{K} \frac{B_{i,j}}{N} |A_{i,j} - C_{i,j}|$$



Calibration	FedAvg			FedProx			Scaffold			FedDyn			FedNova		
Method	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓
Uncal.	61.34	10.52	3.61	61.88	9.37	3.58	62.08	11.42	3.82	62.39	12.53	3.84	63.01	11.45	3.82
Focal (Lin et al., 2017)	60.59	12.88	3.74	62.07	11.45	3.79	61.39	13.28	4.88	60.21	11.22	3.81	61.15	13.21	4.71
LS (Müller et al., 2019)	59.35	15.61	6.28	62.23	14.37	6.14	62.15	11.69	4.05	61.34	15.19	6.58	63.05	16.03	6.89
BS (Brier, 1950)	61.20	11.32	3.80	60.43	10.15	3.61	60.39	10.92	3.80	62.17	13.81	4.91	60.44	11.39	4.06
MMCE (Kumar et al., 2018)	60.00	13.41	4.93	60.11	11.32	3.77	61.03	11.37	3.83	61.35	12.93	3.95	61.28	12.55	3.90
FLSD (Mukhoti et al., 2020)	58.71	11.51	3.92	59.28	10.66	3.81	60.49	13.61	5.15	60.94	11.83	4.04	60.49	13.81	4.88
MbLS (Liu et al., 2021)	59.62	9.42	3.55	60.39	11.37	3.84	61.38	11.99	4.07	62.30	14.95	6.50	62.93	13.66	4.69
DCA (Liang et al., 2020)	61.24	7.69	3.24	61.93	8.74	3.40	62.11	9.25	3.55	62.93	10.17	3.75	63.15	9.27	3.55
NUCFL (DCA+COS)	61.88	6.21	3.11	62.38	8.15	3.35	62.17	8.88	3.50	62.81	9.29	3.54	63.24	8.85	3.51
NUCFL (DCA+L-CKA)	62.05	**6.14**	**3.07**	62.31	**8.04**	**3.30**	62.25	**8.41**	**3.45**	62.94	**9.14**	**3.52**	63.27	8.52	3.43
NUCFL (DCA+RBF-CKA)	61.59	6.19	3.11	61.89	8.17	3.35	61.94	8.52	**3.45**	62.84	9.21	3.55	63.17	**8.01**	**3.30**
MDCA (Hebbalaguppe et al., 2022b)	61.03	7.71	3.29	62.00	8.21	3.37	62.23	9.04	3.51	62.84	10.24	3.72	63.29	10.00	3.71
NUCFL (MDCA+COS)	62.00	6.38	3.14	61.93	7.94	3.29	62.17	8.31	**3.40**	62.91	9.33	3.56	63.14	9.16	3.58
NUCFL (MDCA+L-CKA)	62.17	6.25	3.11	62.03	**7.88**	**3.25**	62.22	**8.30**	**3.40**	62.88	**9.19**	**3.53**	63.14	9.03	3.51
NUCFL (MDCA+RBF-CKA)	61.54	**6.20**	**3.09**	61.79	8.02	3.29	62.15	8.42	3.45	62.65	9.24	3.55	63.22	**8.59**	**3.47**

Table 1: Accuracy (%), calibration measures ECE (%), and SCE (%) of various federated optimization methods with different calibration methods under non-IID scenario on the CIFAR-100 dataset. Values in boldface



Calibration	FedAvg			FedProx			Scaffold			FedDyn			FedNova		
Method	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓
Uncal.	90.95	4.17	1.66	91.45	4.61	1.95	91.28	5.02	2.51	92.75	5.24	2.60	92.22	4.92	2.44
Focal (Lin et al., 2017)	90.63	4.77	1.96	90.00	5.39	2.77	91.13	5.38	2.72	91.15	5.84	2.91	91.13	5.39	2.63
LS (Müller et al., 2019)	91.07	3.75	1.62	90.37	4.52	1.90	91.33	5.18	2.64	93.08	5.19	2.63	92.04	4.95	2.46
BS (Brier, 1950)	91.48	5.19	2.65	90.98	5.42	2.80	88.72	4.19	1.81	91.39	5.37	2.62	90.38	4.65	1.92
MMCE (Kumar et al., 2018)	90.22	4.85	2.30	90.12	4.01	1.79	91.44	5.07	2.55	92.11	4.93	2.49	91.35	5.08	2.49
FLSD (Mukhoti et al., 2020)	90.02	4.93	2.95	90.39	4.99	2.04	92.88	5.19	2.58	91.17	5.61	2.84	91.22	5.23	2.63
MbLS (Liu et al., 2021)	90.62	4.27	1.99	91.49	4.79	1.94	92.87	5.39	2.60	92.06	5.44	2.75	90.39	4.61	1.92
DCA (Liang et al., 2020)	91.84	3.61	1.52	92.03	4.25	1.61	92.04	4.44	1.82	93.08	4.61	1.92	92.37	4.31	1.71
NUCFL (DCA+COS)	91.77	3.52	1.49	91.95	3.61	1.35	92.42	4.39	1.77	93.22	**4.20**	**1.65**	92.25	4.13	1.68
NUCFL (DCA+L-CKA)	91.63	3.52	1.47	92.10	**3.60**	**1.30**	92.19	**4.09**	**1.60**	93.45	4.22	**1.65**	92.33	**4.02**	**1.54**
NUCFL (DCA+RBF-CKA)	91.74	**3.49**	**1.40**	91.99	3.77	1.35	91.74	4.15	1.62	92.95	4.34	1.75	92.41	4.11	1.68
MDCA (Hebbalaguppe et al., 2022b)	91.64	3.75	1.53	92.17	4.42	2.05	92.95	4.61	1.90	93.19	4.71	1.93	92.05	4.62	1.82
NUCFL (MDCA+COS)	91.29	3.61	1.44	91.95	3.95	1.77	93.07	**4.14**	**1.62**	92.88	4.41	1.80	92.45	4.32	1.70
NUCFL (MDCA+L-CKA)	91.97	**3.28**	**1.28**	92.20	**3.88**	**1.51**	92.77	4.20	1.80	93.11	**4.31**	**1.75**	91.99	4.28	1.67
NUCFL (MDCA+RBF-CKA)	91.35	3.56	1.40	92.21	4.00	1.77	93.04	4.19	1.79	93.04	4.40	1.80	92.39	**4.17**	**1.67**

Table 2: Average performance of each algorithm under non-IID scenario on the FEMNIST dataset. Complete results with standard deviation are included in the Appendix.

utilize three similarity measurements–cosine similarity (COS), linear centered kernel alignment (L-CKA), and RBF-CKA

Experiments

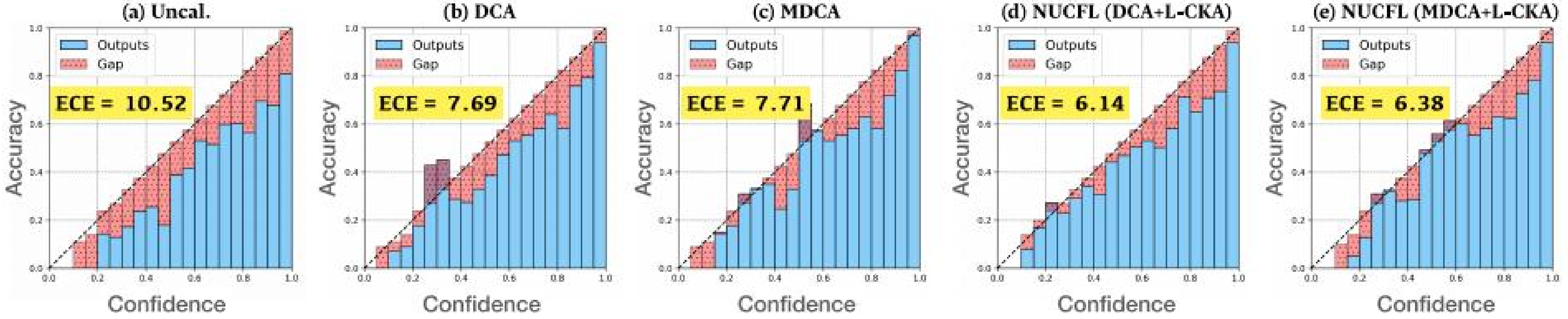


Figure 3: Reliability diagrams for non-IID FedAvg with different calibration methods using the CIFAR-100 dataset. The lower ECE and smaller gap (red region) show the effectiveness of our method.

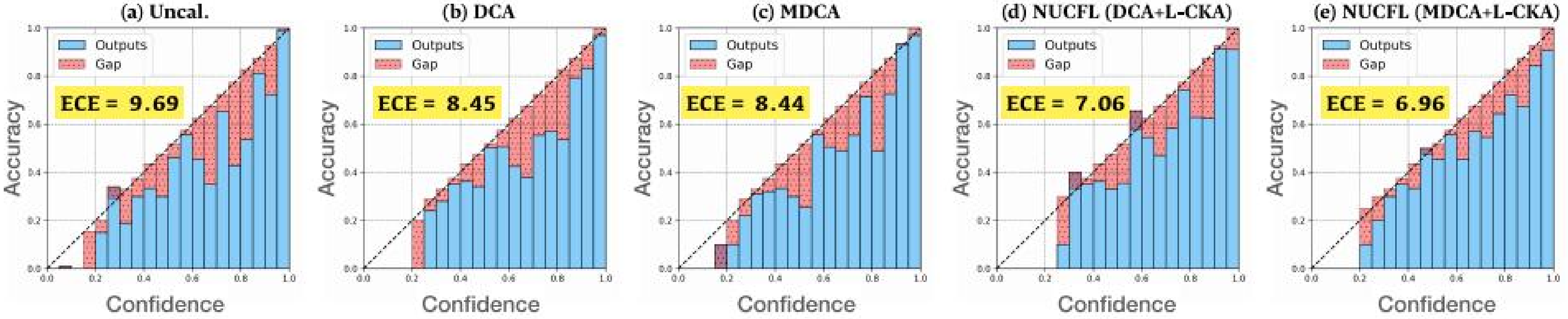
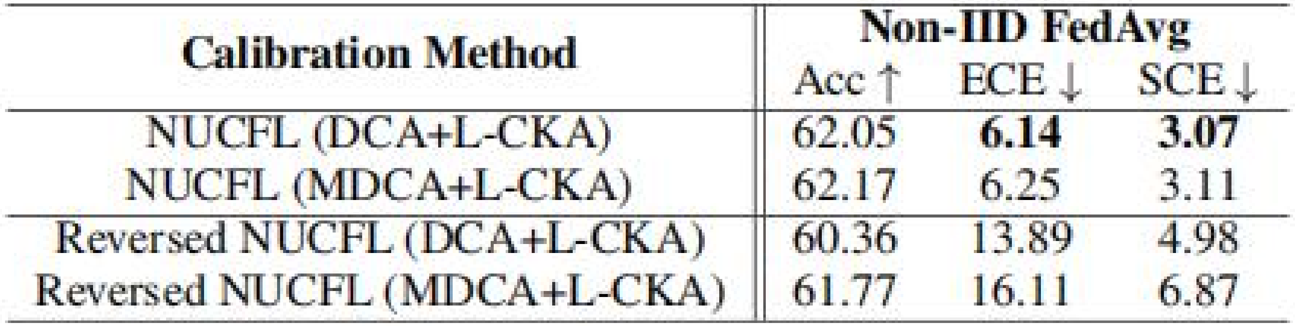


Figure 4: Reliability diagrams for non-IID FedAvg using the CIFAR-10 dataset.

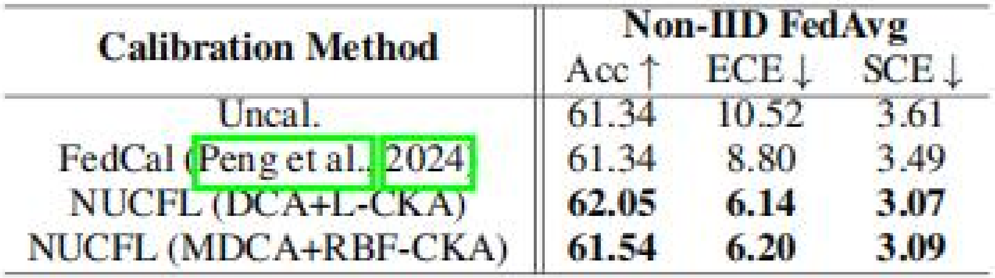
Experiments



Calibration Method	Non-IID FedAvg		
	Acc ↑	ECE ↓	SCE ↓
NUCFL (DCA+L-CKA)	62.05	**6.14**	**3.07**
NUCFL (MDCA+L-CKA)	62.17	6.25	3.11
Reversed NUCFL (DCA+L-CKA)	60.36	13.89	4.98
Reversed NUCFL (MDCA+L-CKA)	61.77	16.11	6.87

Table 3: Destructive experiment for our method.

$$\beta_m = sim(\delta^{(t-1)}, \delta_m^{(t,e)})^{-1}$$



Calibration Method	Non-IID FedAvg		
	Acc ↑	ECE ↓	SCE ↓
Uncal.	61.34	10.52	3.61
FedCal (Peng et al., 2024)	61.34	8.80	3.49
NUCFL (DCA+L-CKA)	**62.05**	**6.14**	**3.07**
NUCFL (MDCA+RBF-CKA)	**61.54**	**6.20**	**3.09**

Table 4: Comparison with FedCal.

- While using a scaler aggregated from local clients can reduce global ECE, it may neglects the interactions between global and local calibration needs.

Experiments

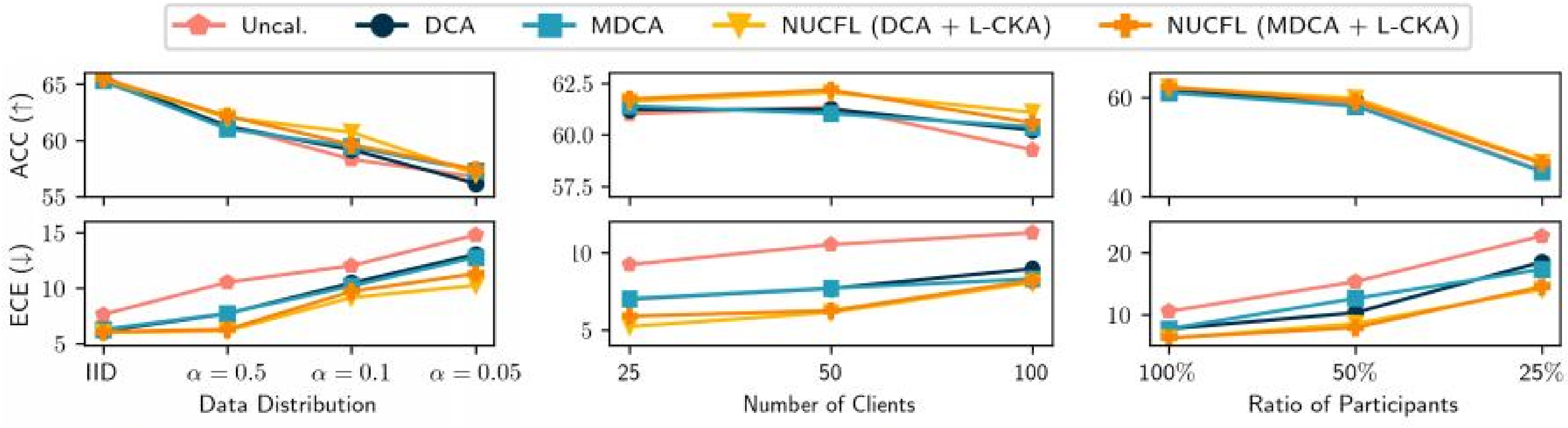
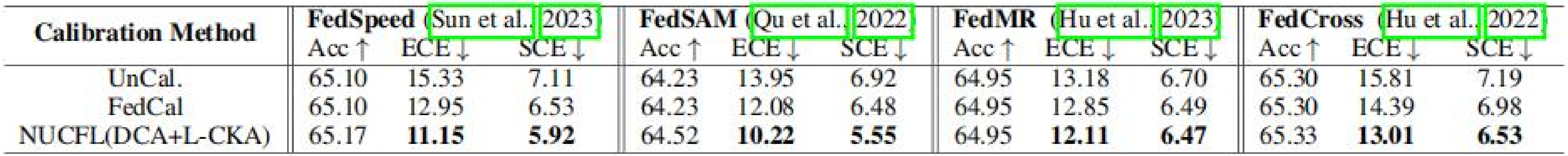


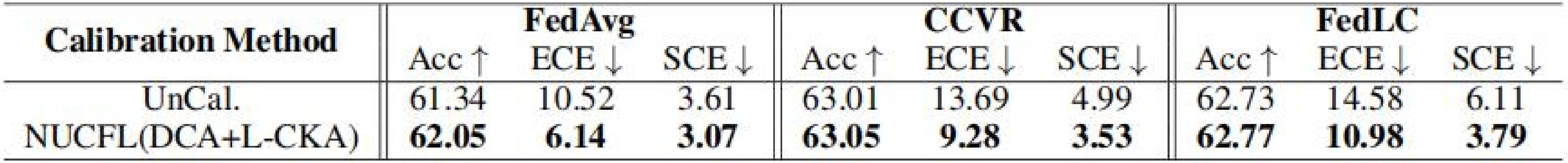
Figure 5: Comparison of confidence calibration across different FL settings using the CIFAR-100 dataset.

Experiments



Calibration Method	FedSpeed (Sun et al., 2023)			FedSAM (Qu et al., 2022)			FedMR (Hu et al., 2023)			FedCross (Hu et al., 2022)		
	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓
UnCal.	65.10	15.33	7.11	64.23	13.95	6.92	64.95	13.18	6.70	65.30	15.81	7.19
FedCal	65.10	12.95	6.53	64.23	12.08	6.48	64.95	12.85	6.49	65.30	14.39	6.98
NUCFL(DCA+L-CKA)	65.17	**11.15**	**5.92**	64.52	**10.22**	**5.55**	64.95	**12.11**	**6.47**	65.33	**13.01**	**6.53**

Table 25: Comparison using other FL optimization methods shows that our method can adapt to any FL algorithm and improve calibration error.



Calibration Method	FedAvg			CCVR			FedLC		
	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓	Acc ↑	ECE ↓	SCE ↓
UnCal.	61.34	10.52	3.61	63.01	13.69	4.99	62.73	14.58	6.11
NUCFL(DCA+L-CKA)	**62.05**	**6.14**	**3.07**	**63.05**	**9.28**	**3.53**	**62.77**	**10.98**	**3.79**

Table 27: Calibration performance of FL algorithms incorporating "calibration."

Thanks