Fully Test-time Adaptation for Tabular Data

Zhi Zhou*, Kun-Yang Yu*, Lan-Zhe Guo ✉, Yu-Feng Li ✉
{zhouz, yuky, guolz, liyf}@lamda.nju.edu.cn
* These authors contributed equally ✉ Corresponding author

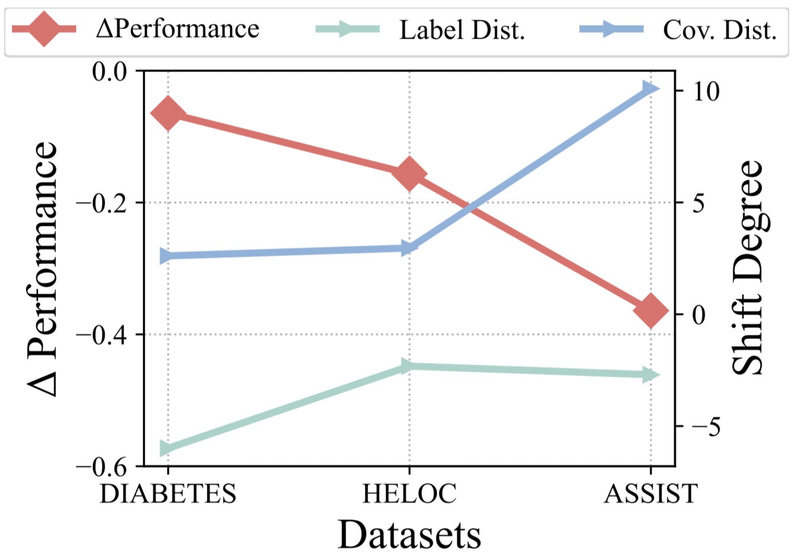
Learning And Mining from DatA

TL; DR We design a fully test-time adaptation for tabular data that addresses covariate and label distribution shifts.

Problem Setting

Distribution shifts in testing data render tabular models ineffective. **Test-time adaptation offers a potential solution, but …**

Observation 1: Covariate distribution and label distribution shifts in tabular data hinder the performance of FTTA methods.



▲ Figure 1: The label and covariate distribution shifts between training and testing in tabular data degrade the model performance. The shift degree is taken logarithm for aesthetic purposes.

Observation 2: Typical augmentation used in test-time adaptation is ineffective for tabular data.

Method	DIABETE	HELOC	ASSIST
Non-Adaptation	**60.82 ± 0.22**	**54.37 ± 5.35**	**55.86 ± 3.81**
$\sigma = 0.2$	60.46 ± 0.20	46.40 ± 3.08	54.89 ± 1.88
$\sigma = 0.4$	59.18 ± 0.42	43.36 ± 0.25	54.86 ± 3.00
$\sigma = 0.6$	57.73 ± 0.64	43.06 ± 0.07	54.51 ± 2.26
$\sigma = 0.8$	56.19 ± 0.83	43.07 ± 0.03	53.79 ± 3.80
$\sigma = 1.0$	54.74 ± 0.77	43.09 ± 0.01	54.23 ± 3.56

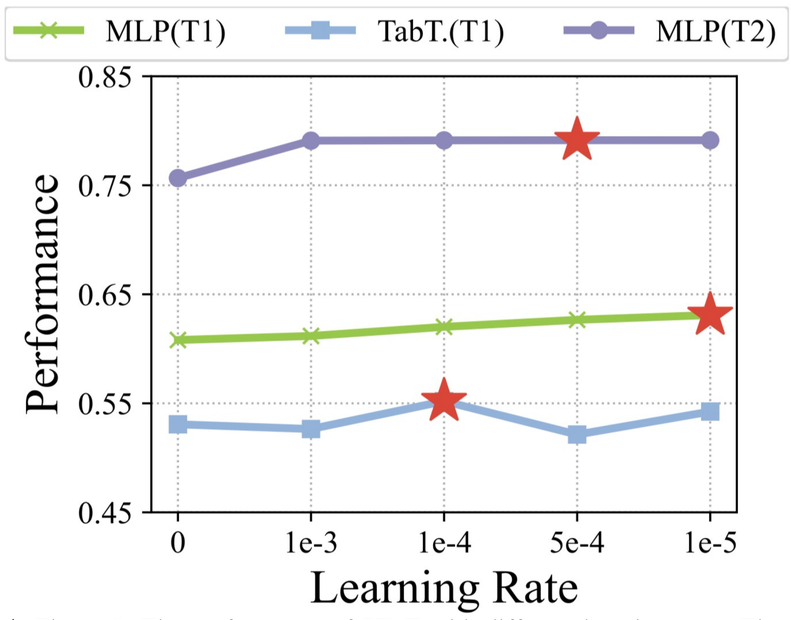
▲ Table 1: Performance of the non-adaptation baseline and CoTTA method with different augmentation strengths σ using the MLP model. The best performance is in bold.

Observation 3: Existing FTTA methods degrades when dealing with tabular data

Method	DIABETE	HELOC	ASSIST
Non-Adaptation	60.82 ± 0.22	54.37 ± 5.35	55.86 ± 3.81
Optimize Parameters	61.34 ± 0.33	54.35 ± 5.38	50.87 ± 0.32
Optimize Predictions	61.47 ± 0.35	43.10 ± 0.00	45.12 ± 0.18

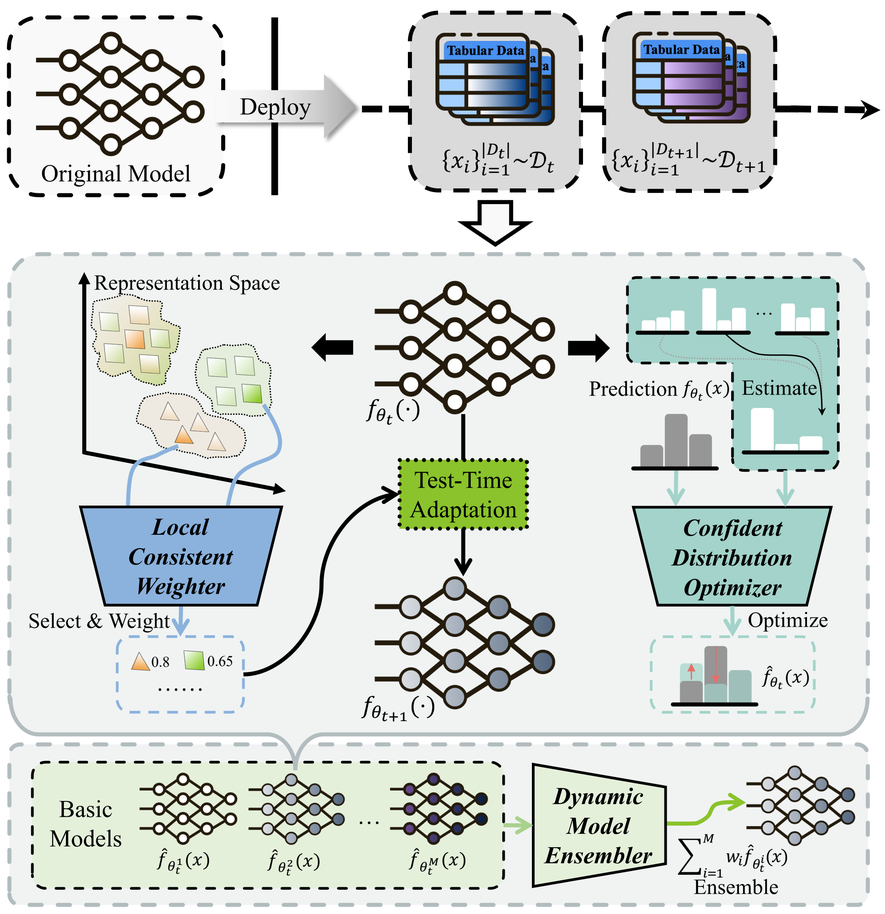
▲ Table 2: Performance of the non-adaptation baseline and two representative FTTA methods using an MLP backbone model. Degraded performance is underlined.

Observation 4: Adaptation is sensitive to both tasks and models for tabular data.



▲ Figure 2: The performance of FTAT with different learning rates. The optimal value differs across backbones and tasks. A red star marks the highest point of each line.

FTAT Method



▲ Figure 3: The overall illustration of FTAT

Confident Distribution Optimizer

Optimize model predictions using the equation:

$$\widehat{f}_{\theta_{t+1}}(x_k) = \frac{f_{\theta_{t+1}}(x_k) \circ \widehat{P}_t}{P_0}$$

where P_0 is training label distribution and $\widehat{P}_t$ is estimated label distribution at timestamp t.

1. Biased estimation $\widetilde{P}_t$ is computed on low-entropy samples: $\widetilde{P}_t = \frac{\sum_{i=1}^{|D_t|} \mathbb{I}\left[\text{Entropy}\left(\hat{f}_{\theta_t}(x_i)\right) < \epsilon\right] \cdot \hat{f}_{\theta_t}(x_i)}{\sum_{i=1}^{|D_t|} \mathbb{I}\left[\text{Entropy}\left(\hat{f}_{\theta_t}(x_i)\right) < \epsilon\right]}$
2. Debias using covariate matrix, where the k-th row is defined as: $\frac{\sum_{i=1}^{|D_t|} \mathbb{I}\left[\arg\max_j \hat{f}_{\theta_t}(x_i)_j = k\right] \cdot \hat{f}_{\theta_t}(x_i)}{\sum_{i=1}^{|D_t|} \mathbb{I}\left[\arg\max_j \hat{f}_{\theta_t}(x_i)_j = k\right]}$
3. Estimated $\widehat{P}_t$ is tracked smoothly throughout the whole time: $\widehat{P}_t = \text{Norm}\left(\widehat{P}_{t-1} - \alpha \cdot \widehat{C}_t^{-1} \widetilde{P}_t\right)$

Local Consistent Weighter

The test-time adaptation is optimized based on the entropy minimization objective with sample weight $\mathcal{W}(x_k, D_t, \theta_t)$:

$$\underbrace{\left[\max \hat{f}_{\theta_t}(x_k) - \min \hat{f}_{\theta_t}(x_k)\right]}_{A1} \cdot \underbrace{\mathcal{I}(x_k, D_t, \theta_t)}_{A2}$$

where $A1$ indicates the margin of prediction and $A2$ is the consistency indicator of prediction.

1. The neighbor set of the sample is defined as follows: $N(x_k, D_t) = \left\{x | \text{Dist}(x, x_k) < \overline{\text{Dist}_t}, x \in D_t\right\}$
2. The threshold is the average distance between samples in the data batch at each timestamp t: $\overline{\text{Dist}_t} = \frac{2}{|D_t|(|D_t|-1)} \sum_{i=1}^{|D_t|} \sum_{j=i+1}^{|D_t|} \text{Dist}(x_i, x_j)$
3. The consistency is defined as the consistency between its prediction and the predictions of its neighbors: $\mathcal{I}(x_k, D_t, \theta_t) = \begin{cases} 1, & \left\| f_{\theta_t}(x_k) - \frac{\sum_{x \in N(x_k, D_t)} f_{\theta_t}(x)}{|N(x_k, D_t)|} \right\| < \beta, \\ 0, & Otherwise. \end{cases}$

Dynamic Model Ensembler

The final predictions are defined as the dynamic ensemble of adapted models using different learning rates:

$$\sum_{i=1}^{M} w_i \cdot \widehat{f}_{\theta_t^i}(x)$$

1. The weight is proportional to the loss $R_t^i(D_t)$: $w_i \propto 1 - R_t^i(D_t)$
2. The weight satisfied the constraint: $\sum_{i=1}^{M} w_i = 1$

Experiments

Method	MLP			TabTransformer			FT-Transformer		
	Acc.	Balanced Acc.	F1	Acc.	Balanced Acc.	F1	Acc.	Balanced Acc.	F1
Non-Adaptation	62.45	64.61	60.59	60.86	63.08	58.32	59.69	62.49	54.29
TENT	58.43	61.63	50.97	58.32	61.40	51.73	51.61	55.41	36.34
EATA	61.43	63.69	60.11	60.33	62.36	60.09	56.04	58.95	44.62
LAME	59.48	62.32	58.47	59.15	62.50	58.39	58.90	61.98	51.86
CoTTA	61.59	63.78	60.57	60.37	62.82	59.75	59.64	62.43	53.41
ODS	59.18	62.15	57.83	59.22	62.02	58.46	59.05	61.70	51.41
SAR	61.16	63.49	59.18	60.30	62.77	59.72	59.27	62.11	57.04
FTAT	**66.77**	**64.96**	**72.00**	**66.14**	**64.40**	**69.03**	**64.01**	**62.54**	**69.56**

▲ Table 3: Average performance of each method on 6 datasets using different backbones.

Method	HELOC			ANES			Health Ins.		
	Acc.	Balanced Acc.	F1	Acc.	Balanced Acc.	F1	Acc.	Balanced Acc.	F1
Non-Adaptation	54.37 ± 5.35	58.25 ± 3.56	40.02 ± 16.8	79.11 ± 0.31	75.66 ± 0.46	**84.24 ± 0.16**	65.79 ± 0.63	**70.68 ± 0.44**	66.21 ± 0.90
TENT	54.35 ± 5.38	58.24 ± 3.58	39.95 ± 16.9	78.07 ± 0.35	74.09 ± 0.65	83.76 ± 0.13	64.30 ± 0.70	69.79 ± 0.47	63.87 ± 1.06
EATA	54.37 ± 5.35	58.25 ± 3.56	40.02 ± 16.8	78.13 ± 0.30	74.20 ± 0.59	83.79 ± 0.10	65.78 ± 0.63	70.68 ± 0.44	66.21 ± 0.90
LAME	43.10 ± 0.00	50.00 ± 0.00	30.10 ± 0.00	63.50 ± 0.00	54.60 ± 0.00	46.80 ± 0.00	63.44 ± 1.69	69.14 ± 1.09	62.61 ± 2.69
CoTTA	54.36 ± 5.35	58.25 ± 3.56	40.03 ± 16.8	78.13 ± 0.30	74.20 ± 0.59	83.79 ± 0.10	65.79 ± 0.63	**70.68 ± 0.44**	66.21 ± 0.90
ODS	43.10 ± 0.00	50.00 ± 0.00	30.10 ± 0.00	63.50 ± 0.00	54.60 ± 0.00	46.80 ± 0.00	63.45 ± 1.68	69.14 ± 1.07	62.62 ± 2.68
SAR	52.32 ± 6.05	56.74 ± 3.99	33.16 ± 19.0	78.13 ± 0.30	74.20 ± 0.59	83.79 ± 0.10	65.79 ± 0.63	**70.68 ± 0.44**	66.21 ± 0.90
FTAT	**64.09 ± 1.14**	**63.64 ± 0.93**	**67.80 ± 2.71**	**80.09 ± 0.23**	**79.12 ± 0.20**	83.42 ± 0.25	**72.42 ± 0.20**	65.30 ± 0.15	**80.83 ± 0.23**
Method	ASSIST			DIABETE			Hypertension		
	Acc.	Balanced Acc.	F1	Acc.	Balanced Acc.	F1	Acc.	Balanced Acc.	F1
Non-Adaptation	55.86 ± 3.81	60.81 ± 3.37	66.42 ± 1.86	60.81 ± 0.21	60.59 ± 0.24	51.18 ± 1.69	58.76 ± 1.68	**61.69 ± 0.95**	55.46 ± 4.03
TENT	50.87 ± 0.32	56.41 ± 0.29	63.99 ± 0.15	61.34 ± 0.33	61.15 ± 0.34	53.75 ± 1.01	41.67 ± 0.08	50.07 ± 0.05	0.49 ± 0.36
EATA	55.86 ± 0.18	60.81 ± 0.16	66.42 ± 0.08	61.36 ± 0.30	61.16 ± 0.31	53.68 ± 1.09	57.81 ± 2.32	61.19 ± 1.38	52.87 ± 5.82
LAME	45.12 ± 0.18	51.30 ± 0.18	61.40 ± 0.18	61.47 ± 0.35	61.30 ± 0.37	54.67 ± 1.45	58.63 ± 1.60	61.64 ± 0.92	55.12 ± 3.84
CoTTA	55.86 ± 0.18	60.81 ± 0.16	66.42 ± 0.08	61.39 ± 0.29	61.20 ± 0.30	53.82 ± 1.05	58.76 ± 1.68	**61.69 ± 0.95**	55.46 ± 4.03
ODS	45.12 ± 0.18	51.30 ± 0.18	61.40 ± 0.18	61.47 ± 0.35	61.30 ± 0.37	54.69 ± 1.43	57.12 ± 1.46	60.80 ± 0.93	51.41 ± 3.43
SAR	55.86 ± 0.18	60.81 ± 0.16	66.42 ± 0.08	61.38 ± 0.30	61.19 ± 0.30	53.98 ± 0.93	58.21 ± 1.51	61.50 ± 0.77	53.81 ± 4.05
FTAT	**60.17 ± 2.87**	**63.79 ± 2.18**	**66.92 ± 1.20**	**61.66 ± 0.30**	**61.54 ± 0.28**	**59.27 ± 0.96**	**62.20 ± 0.94**	56.36 ± 1.62	**73.77 ± 0.13**

▲ Table 4: Performance of FTAT approach and comparison methods on 6 datasets using MLP.

RQ1: Does FTAT method outperform comparisons across different datasets using different backbones?

RQ2: Does each component contribute to performance in the FTAT method?

RQ3: Does the FTAT method accurately estimate the label distributions?

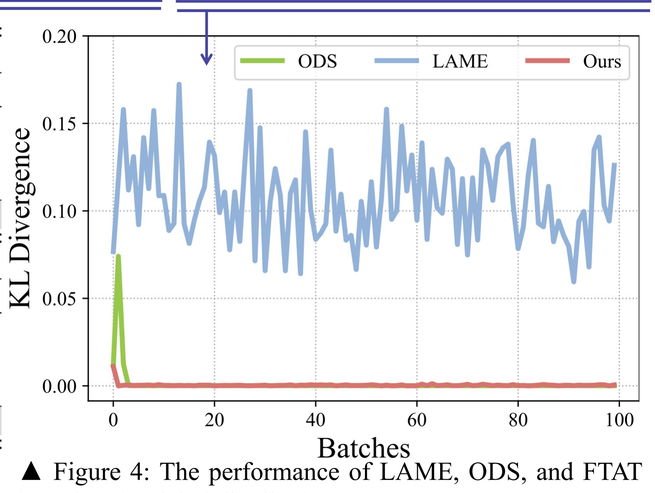
DIABETE

Method	Acc.	Balanced Acc.	F1
Non-Adaptation	60.81 ± 0.21	60.59 ± 0.23	51.18 ± 1.69
FTAT w/o CDO	60.85 ± 0.22	60.61 ± 0.24	51.26 ± 1.69
FTAT w/o LCW	61.43 ± 0.16	61.28 ± 0.20	55.61 ± 1.67
FTAT	**61.66 ± 0.30**	**61.54 ± 0.28**	**59.27 ± 0.96**

HELOC

Method	Acc.	Balanced Acc.	F1
Non-Adaptation	54.37 ± 5.35	58.25 ± 3.56	40.02 ± 16.8
FTAT w/o CDO	62.73 ± 0.23	62.55 ± 0.52	66.45 ± 0.91
FTAT w/o LCW	62.52 ± 0.86	61.73 ± 1.06	67.07 ± 1.20
FTAT	**64.09 ± 1.14**	**63.64 ± 0.93**	**67.80 ± 2.71**

▲ Table 5: Ablation study. The performance of the FTAT approach using MLP backbone when removing different components.



▲ Figure 4: The performance of LAME, ODS, and FTAT in estimating label distribution.

✓ If you are interested in this paper, feel free to contact Zhi Zhou and Kun-Yang Yu (zhouz@lamda.nju.edu.cn, yuky@lamda.nju.edu.cn).
✓ To obtain more details of our paper, please visit the project homepage (https://wnjxyk.github.io/FTTA).
✓ This research was supported by the NSFC (Grant No. 624B2068, 62176118, and 62306133), the Key Program of Jiangsu Science Foundation (BK20243012), and the Fundamental Research Funds for the Central Universities (022114380023).