A Appendix

A.1 Effect of Number of Quantiles

We examine the effect of the number of quantiles on the performance of TDTransformer. Figure 5 shows the performance comparison. When decreasing the number of quantiles from 64 to 8, we observe a performance degradation. When increasing the number of quantiles from 64 to 256, we do not find a large performance gain.

PLE does not break the continuity of the original numerical values. The number of quantiles determines the granular level of dividing a continuous range into different segments. If the number is equal to 1, PLE is similar to min-max normalization. The difference is that PLE maps scalars to the range $[-1, 1]$ while min-max normalization maps scalars to the range $[0, 1]$. At this time, PLE does not utilize the high dimensional vector form to indicate the statistical distribution information. If the number is infinitely large, on the other hand, the segmentation of a continuous range is highly affected by noise within data. Even a small noise level will lead to different quantiles.

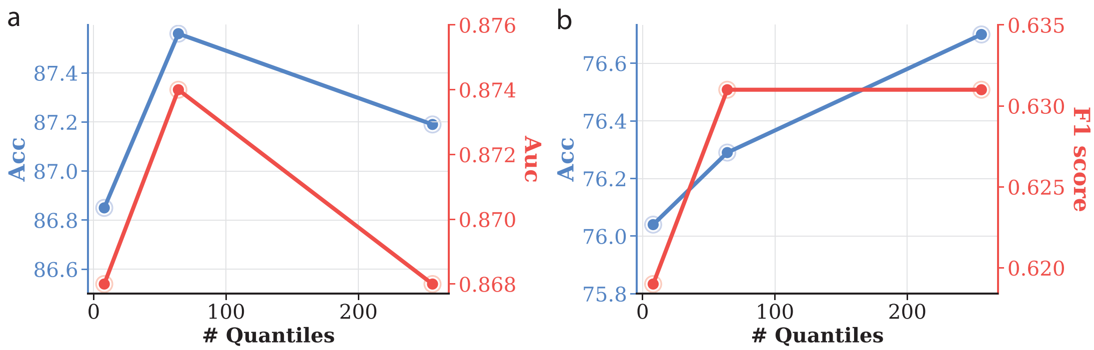

Figure 5: The effect of the number of quantiles on the model performance. Metrics are computed over tables that contain numerical columns. We examine the number of quantiles in $\{8, 64, 256\}$. (a) Binary classification task. (b) Multiclass classification task.

A.2 Effect of Model Complexity

Figure 6 shows the effect of the model complexity on the model performance in both binary classification tasks and multiclass classification tasks. As the number of trainable parameters increases, the average test accuracy in the binary classification task decreases whereas that in the multiclass classification task increases. Hence, in relatively easy binary classification tasks, decreasing model complexity can achieve the boost in both performance and efficiency. In more challenging multiclass classification tasks, increasing computation efficiency has the cost of performance degradation.

A.3 Effect of Backbone Models

In addition to the gated transformer (Wang & Sun, 2022), we examine the performance of the TDTransformer framework using RoBERTa (Liu, 2019) as the backbone model. Table 7 shows the performance comparison for the binary classification task and Table 8 shows the performance comparison for the multiclass classification task. Figure 7 shows the comparison of the average performance. Overall, gated transformer as the backbone model has a similar performance to RoBERTa.

A.4 Dataset Details

We summarize the statistics of tables for the binary classification task in Table 9. The total number of tables for the binary classification is 36. In the OpenML benchmark, tables are categorized into “categorical

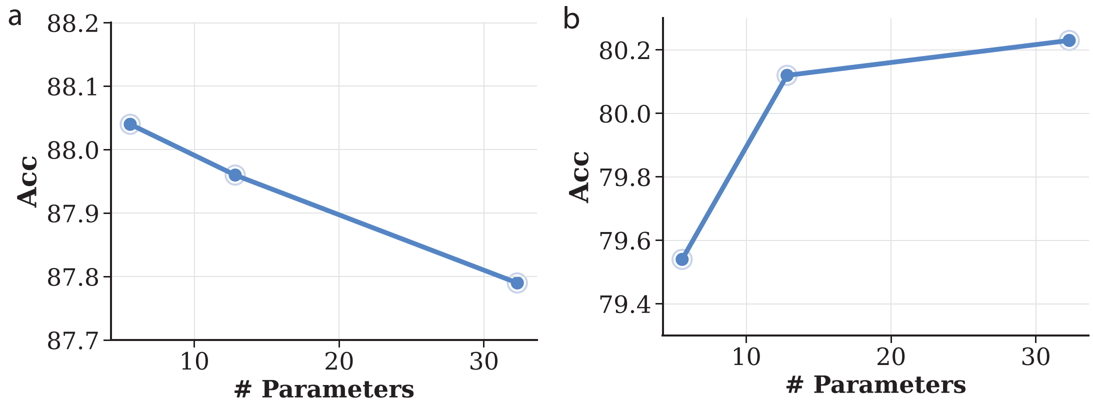

Figure 6: The effect of the number of trainable parameters on the model performance. We choose different hidden dimension sizes in {512, 256, 128}. (a) Binary classification task. (b) Multiclass classification task.

Table 7: Performance comparison of backbone models for the binary classification task.

Method	$\mathcal{S} \cup \mathcal{S}_{\text{num}}$		$\gamma \leq 0.2$		$0.2 < \gamma < 0.8$		$\gamma \geq 0.8$		Avg	
	Acc	Auc	Acc	Auc	Acc	Auc	Acc	Auc	Acc	Auc
Gated Transformer	87.56	0.87	91.67	0.87	83.94	0.88	95.40	0.96	87.79	0.88
RoBERTa	87.57	0.86	91.70	0.85	84.14	0.87	95.49	0.95	87.92	0.87

columns" and binary columns. We refine the categorization by splitting "categorical columns" into binary columns (cell values are True/False or T/F, or 0/1) and categorical columns.

Table 10 shows the statistics of tables for the multiclass classification task. The total number of tables for the multiclass classification task is 40. Same as the binary classification task, we categorize tables into categorical columns, binary columns and numerical columns.

A.5 Implementation Details on Baseline Methods

To ensure approximately the same complexity, we use the hidden dimension of 512 and the model depth of 12 for all transformer-based architectures. The pre-training and fine-tuning processes use the early stopping strategy with a patience of 10. Batch size N_{bs} is 128. The maximum number of training epochs is 200. Except for hyperparameters controlling the model size, we use default hyperparameters in baselines.

A.6 Comparison of Computational Cost

We compare the total number of trainable parameters with deep learning methods. Figure 8 shows the comparison. Different deep learning methods have a similar number of trainable parameters. The maximum variation in the total number of trainable parameters is smaller than 2.4% while the maximum performance variation is larger than 20%.

Table 8: Performance comparison of backbone models for the multiclass classification task.

Method	$\mathcal{S} \cup \mathcal{S}_{\text{num}}$		$\lvert\mathcal{D}\rvert < 2000$		$\lvert\mathcal{D}\rvert \geq 2000$		$\mathfrak{C} < 10$		$\mathfrak{C} \geq 10$		Avg	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Gated Transformer	76.30	0.63	78.68	0.69	81.06	0.70	80.89	0.65	79.00	0.77	80.23	0.70
RoBERTa	76.48	0.61	78.86	0.67	81.11	0.69	80.53	0.63	79.93	0.78	80.32	0.68

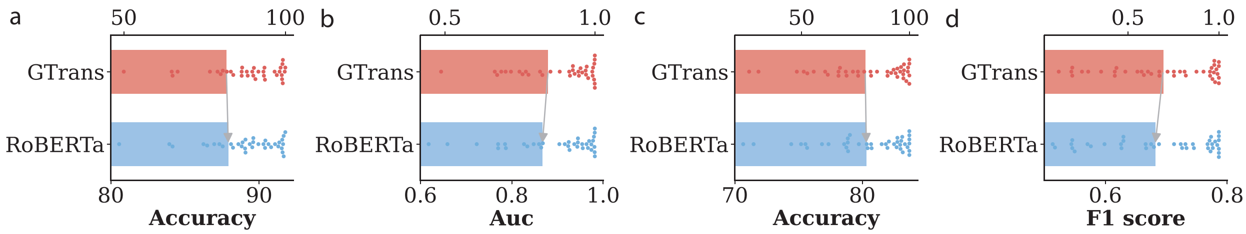

Figure 7: The performance comparison between the backbone model of gated transformer (Wang & Sun, 2022) and RoBERTa (Liu, 2019). Overall, these two backbone models have a similar performance.

Table 9: Statistics of tables for the binary classification task.

Dataset	Spambase	Telco-Customer	Credit	QSar	Arrhythmia	Blood-Transfusion	Tic-Tac-Toe
Size	4,601	7,043	1,000	1,055	452	748	958
# Cat	0	11	11	41	37	0	0
# Bin	0	5	2	0	36	0	0
# Num	57	3	7	0	206	4	9
Dataset	Steel-Plates	Phoneme	WDBC	KC2	Climate	ILPD	PC1
Size	1,941	5,404	569	522	540	583	1,109
# Cat	0	0	0	0	0	1	0
# Bin	0	0	0	0	0	0	0
# Num	33	5	30	21	20	9	21
Dataset	PC4	PC3	Scene	Sick	Churn	Ailerons	BankNote
Size	1,458	1,563	2,407	3,772	5,000	7,129	1,372
# Cat	0	0	0	3	2	0	0
# Bin	0	0	5	19	2	0	0
# Num	37	37	294	7	16	5	4
Dataset	Wilt	Satellite	Pollen	BankMarket	JapaneseVowels	MC1	Kin8NM
Size	4,839	5,100	3,848	4,521	9,961	9,466	8,192
# Cat	0	0	0	6	0	0	0
# Bin	0	0	0	3	0	0	0
# Num	5	36	5	7	14	38	8
Dataset	Karhunen	Elevators	EyeState	Mozilla	JM1	BankMarketing	ClickPredict
Size	2,000	9,517	14,980	15,545	10,885	45,211	39,948
# Cat	0	0	0	0	0	6	0
# Bin	0	0	0	0	0	3	0
# Num	63	6	14	5	21	7	9
Dataset	Mushroom						
Size	8,124						
# Cat	21						
# Bin	1						
# Num	0						

We compare the running time of TDTransformer with tree-based methods. Similar to earlier works comparing tree-based methods and deep learning methods (Grinsztajn et al., 2022; Borisov et al., 2022; Zabërgja et al., 2024), the running time of TDTransformer is longer than tree-based methods XGBoost (Chen & Guestrin, 2016) and CatBoost (Prokhorenkova et al., 2018; Dorogush et al., 2018) as shown in Table 11.

Table 10: Statistics of tables for the multiclass classification task.

Dataset	Eucalyptus	CarEval	SolarFlare	Car	Okcupic	Letter	Soybean
Size	736	1,728	1,066	1,728	50,789	20,000	683
# Cat	5	7	12	6	17	0	33
# Bin	0	14	0	0	0	0	2
# Num	14	0	0	0	2	16	0
# Class	5	4	6	4	3	26	19
Dataset	Karhunen	Fourier	Factors	Morphological	PlantsMargin	PlantsShape	PlantsTexture
Size	683	2,000	2,000	2,000	1,600	1,600	1,599
# Cat	35	0	0	0	0	0	0
# Bin	0	0	0	0	0	0	0
# Num	0	76	216	6	64	64	64
# Class	19	10	10	10	100	100	100
Dataset	OptDigits	MiceProtein	Au7-1100	Au4-2500	Baseball	Zernike	SatImage
Size	5,620	1,080	1,100	2,500	1,340	2,000	6,430
# Cat	0	4	4	42	1	0	0
# Bin	0	0	0	0	0	0	0
# Num	64	77	8	58	15	47	36
# Class	10	8	5	3	3	10	6
Dataset	Theorem	Navigation	Abalone	Gesture	Characters	GasDrift	Nursery
Size	6,118	5,456	4,177	9,873	10,218	13,910	12,960
# Cat	0	0	0	0	0	0	8
# Bin	0	0	0	0	0	0	0
# Num	51	24	8	32	7	128	0
# Class	6	4	28	5	10	6	5
Dataset	Kropt	SleepData	CJS	Splice	Cardiotography	Volcanoes-a3	Volcano-d3
Size	28,056	1,024	2,796	3,196	2,126	1,521	9,285
# Cat	6	0	2	60	0	0	0
# Bin	0	0	0	0	0	0	0
# Num	0	2	32	0	35	3	3
# Class	18	4	6	3	3	5	5
Dataset	Volcano-d1	Nursery (VR)	RobotNavigation	ThyroidAllBP	ThyroidAllhyper		
Size	8,753	12,958	5,456	2,800	2,800		
# Cat	0	8	0	20	20		
# Bin	0	0	0	0	0		
# Num	3	0	2	6	6		
# Class	5	4	4	5	5		

Table 11: Running time (in minutes) comparison with tree-based methods. The running time for the binary classification task is averaged over 36 tables. The running time for the multiclass classification task is averaged over 40 tables.

Method	Binary classification			Multiclass classification		
	Time (avg) ↓	Accuracy ↑	Auc ↑	Time (avg) ↓	Accuracy ↑	F1 ↑
TDTransformer	32.45	87.79	0.88	36.67	80.23	0.70
TDTransformer (CTA Pos)	32.04	87.48	0.87	36.53	80.51	0.70
XGBoost	6.92	84.97	0.83	7.48	76.45	0.66
CatBoost	6.94	86.12	0.87	7.28	76.61	0.65

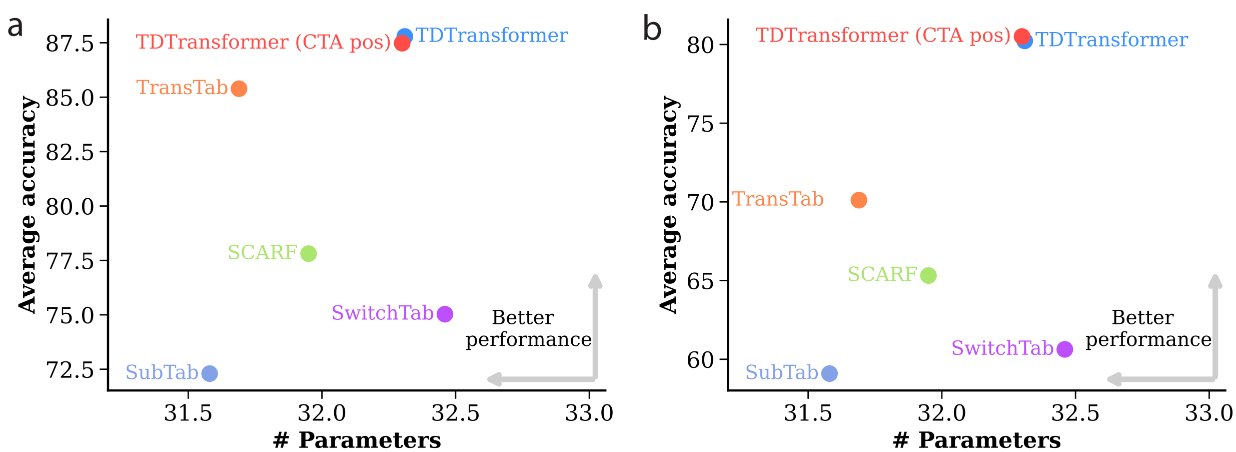

Figure 8: Comparison of the number of trainable parameters with deep learning methods. (a) Binary classification task. (b) Multiclass classification task. The maximum variation in the total number of trainable parameters is within 2.4% of model parameters.