A Appendix

A.1 Hyperparameters for GBDT Models

To evaluate the hyperparameters for gradient boosted decision tree models used in [11], we train 35 models for each dataset to conduct a grid search. In details, we use maximum depth: 4, 5, 6, 7, 8, 9, 10; number of trees for breast-cancer: 2, 4, 6, 8, 10, cod-rna: 10, 20, 30, 40, ijcnn1: 20, 40, 60, 80, and binary mnist: 600, 800, 1000, 1200. Table 10 and 11 reports the model hyperparameters and corresponding test accuracy of trained models which obtain the best validation accuracy. In comparison with the results from Table 3, the hyperparameters used by [11] can train models with accuracy similar to the best one.

Dataset	Trained ε		Tree Num / Depth		
	Chen's	ours	natural	Chen's	ours
breast-cancer	0.30	0.30	4 / 6	4 / 4	2 / 7
cod-rna	0.20	0.03	40 / 10	40 / 4	10 / 10
ijcnn1	0.20	0.02	80 / 5	80 / 10	80 / 10
MNIST 2 vs. 6	0.30	0.30	600 / 4	600 / 8	600 / 9

Table 10: GBDT model hyperparameters with the best validation accuracy in XGBoost.

Dataset	Test ACC (%)			Test FPR (%)		
	natural	Chen's	ours	natural	Chen's	ours
breast-cancer	97.81	96.35	99.27	0.98	0.98	0.98
cod-rna	96.74	87.32	91.08	2.79	4.05	8.71
ijcnn1	97.85	97.24	93.66	1.74	1.53	1.70
MNIST 2 vs. 6	99.70	99.65	99.55	0.39	0.39	0.29

Table 11: Test accuracy of GBDT models with the best validation accuracy in XGBoost.

A.2 Recall for Twitter Spam Models

To evaluate the performance of all 23 models trained to detect Twitter spam, we computed the recall at 1% FPR, 5% FPR, and 10% FPR in Table 12. The models M1, M6, M10, and M16 have the best recall within their cost family.

Classifier Model	**Adaptive Objective**	**Model Quality**		
		1% FPR Recall	5% FPR Recall	10% FPR Recall
Natural	-	0.9974	0.9998	0.9999
C1	-	0.8177	0.9844	0.9999
C2	-	0.7912	0.9250	0.9897
C3	-	0.6928	0.8609	0.8609
M1	$Cost_1$	**0.9612**	**0.9992**	**0.9997**
M2		0.7949	0.9893	0.9973
M3		0.8214	0.9948	0.9981
M4		0.7537	0.9281	0.9689
M5		0.6907	0.9280	0.9840
M6	$Cost_2$	**0.9162**	**0.9948**	**0.9968**
M7		0.7881	0.9901	0.9959
M8		0.6793	0.9220	0.9608
M9		0.6780	0.9016	0.9386
M10	$Cost_3$	**0.9715**	**0.9996**	**0.9999**
M11		0.8671	0.9948	0.9991
M12		0.7484	0.9846	0.9930
M13		0.7753	0.9383	0.9896
M14		0.7473	0.9806	0.9925
M15		0.6728	0.8852	0.9862
M16	$Cost_4$	0.8624	0.9929	**0.9989**
M17		**0.9061**	**0.9946**	0.9973
M18		0.7075	0.9368	0.9749
M19		0.7298	0.9361	0.9703

Table 12: Recall at 1% FPR, 5% FPR, and 10% FPR for all Twitter spam detection models. The best recall numbers highlighted in bold.