

Online Supplementaries

Learning with Confidence: Training Better Classifiers from Soft Labels

Sjoerd de Vries^{1,2*} and Dirk Thierens¹

¹*Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC, Utrecht, The Netherlands.

²Department of Digital Health, University Medical Center Utrecht,
Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands.

1 Question 1: Heat Maps

Shared description for all Figures in Online Supplementary Section 1: Heat map illustrating the performance of various methods with the mentioned base classifier across multiple datasets measured by either AUC on y^G or $\overline{TV\bar{D}}$ on y^{PG} . In the last column, the average of their rankings over all datasets is shown, with lower values indicating better performance. All values were multiplied by 100 to enhance readability. Red cells indicate higher AUC values, while blue cells represent lower values relative to the performance of the PluralityBootstrapClf for each dataset.

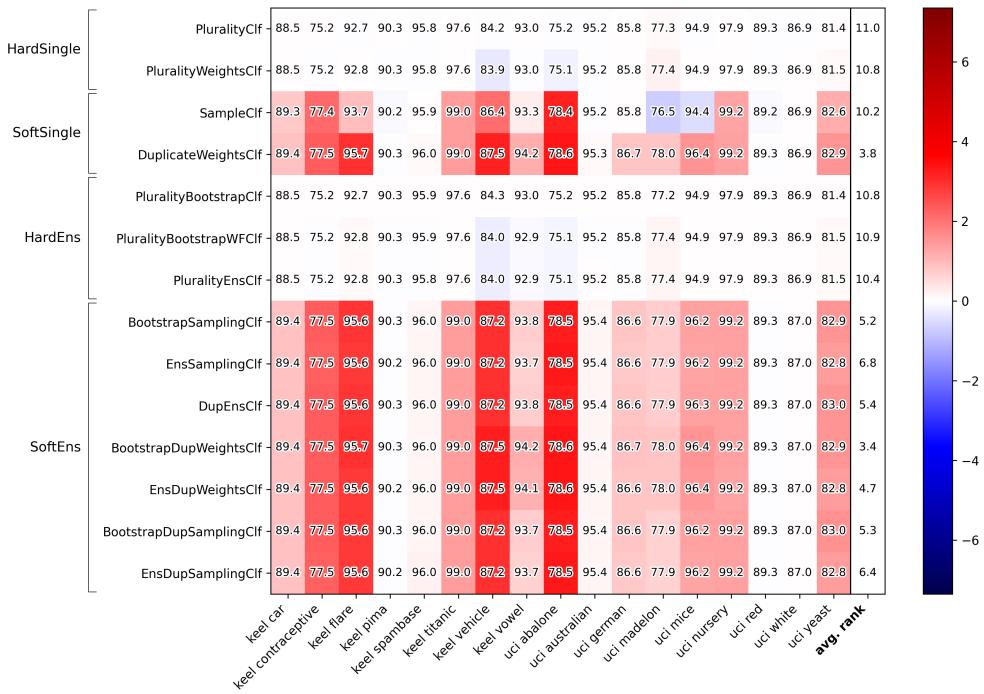


Fig. OS.1: Base classifier: LR. Metric: AUC on y^G .

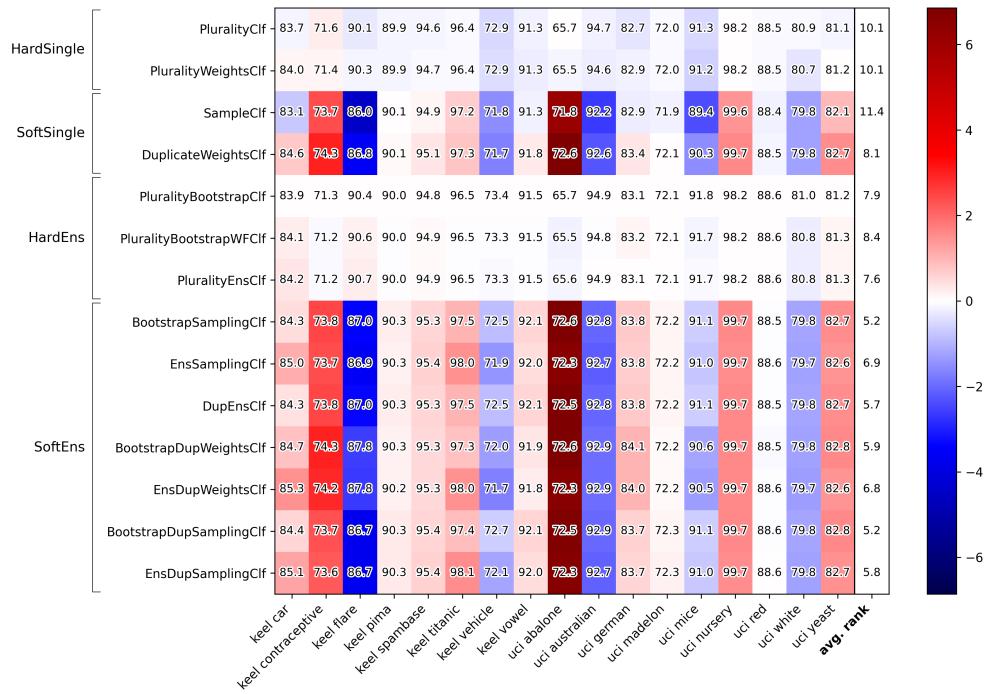


Fig. OS.2: Base classifier: GNB. Metric: AUC on y^G .

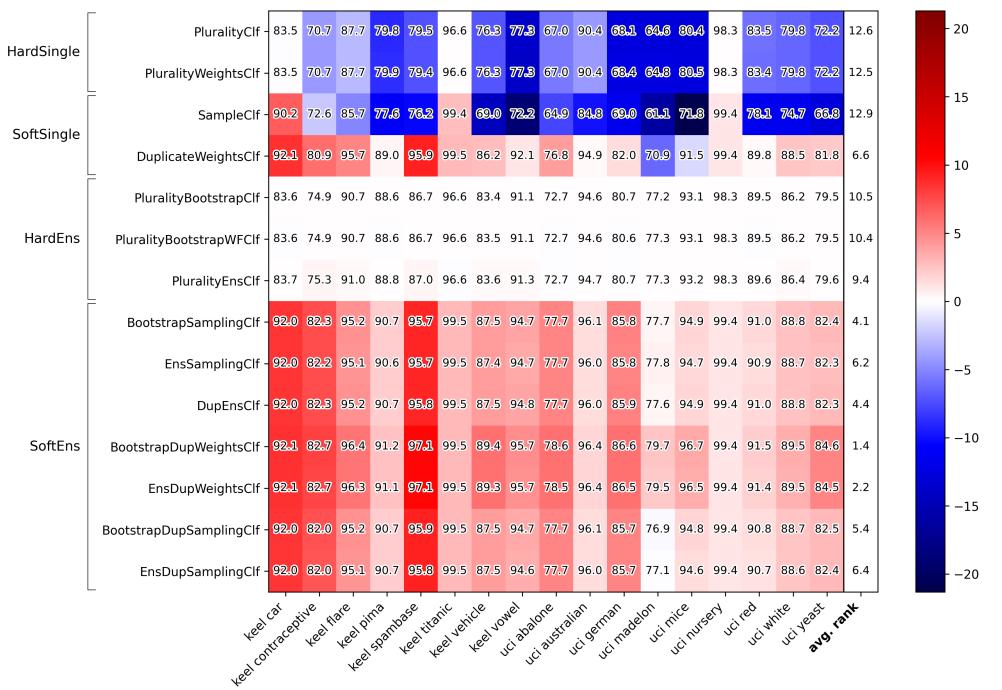


Fig. OS.3: Base classifier: DT. Metric: AUC on y^G .

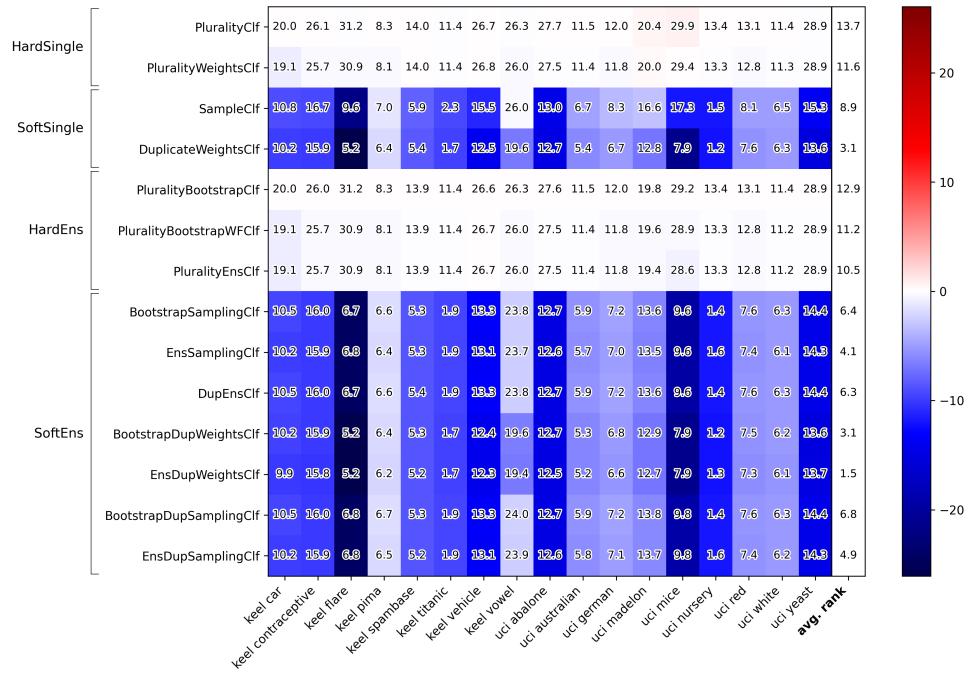


Fig. OS.4: Base classifier: LR. Metric: \overline{TVD} on y^{PG} .

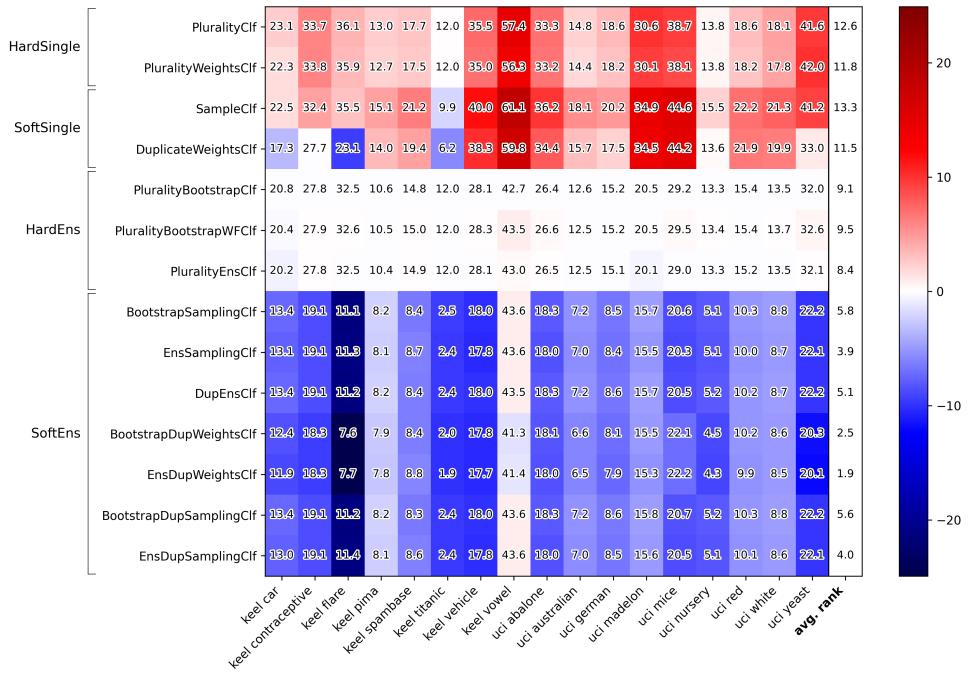


Fig. OS.5: Base classifier: SGD. Metric: \overline{TVD} on y^{PG} .

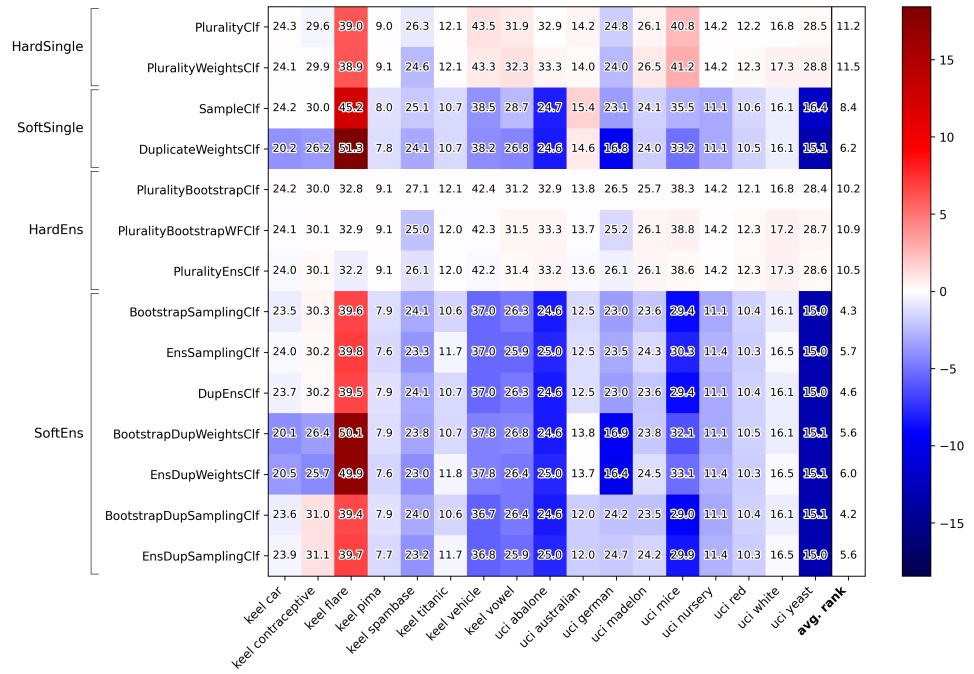


Fig. OS.6: Base classifier: GNB. Metric: \overline{TVD} on y^{PG} .

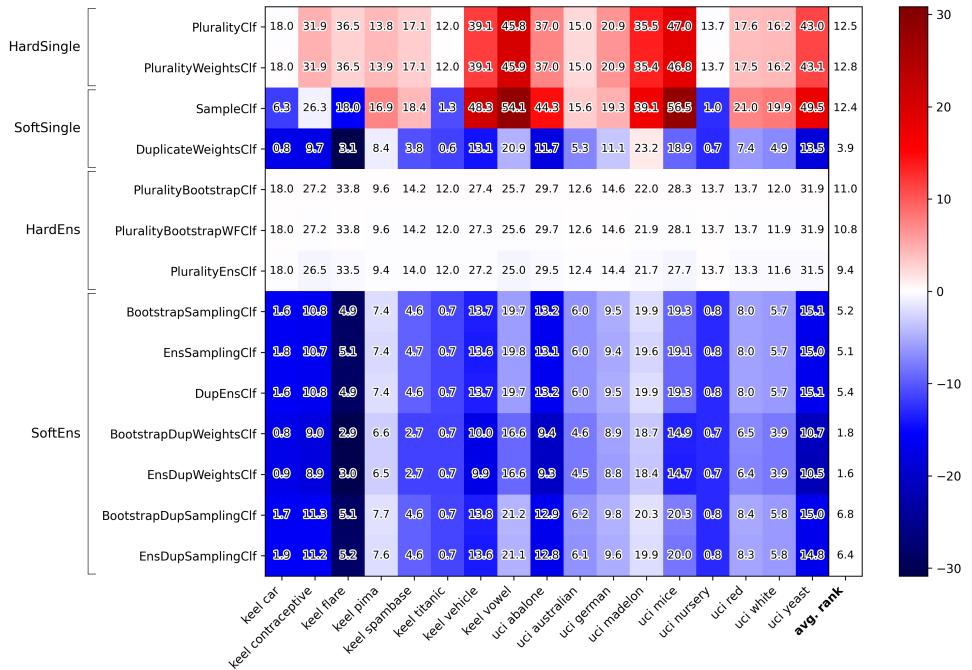


Fig. OS.7: Base classifier: DT. Metric: \overline{TVD} on y^{PG} .

2 Question 2: Effect of Noise

Shared description for all Figures in Online Supplementary Section 2: The effect of six different noise types on method performance with four base classifiers, measured by \overline{TVD} on the partial ground truth test data, across multiple noise levels. Noise levels range from level 0 (noiseless) to level 6 (noise strength 0.3). LR, SGD, GNB and DT were used as base classifiers. RF served as the ground truth model, with the soft labels generated at either the low or high uncertainty level.

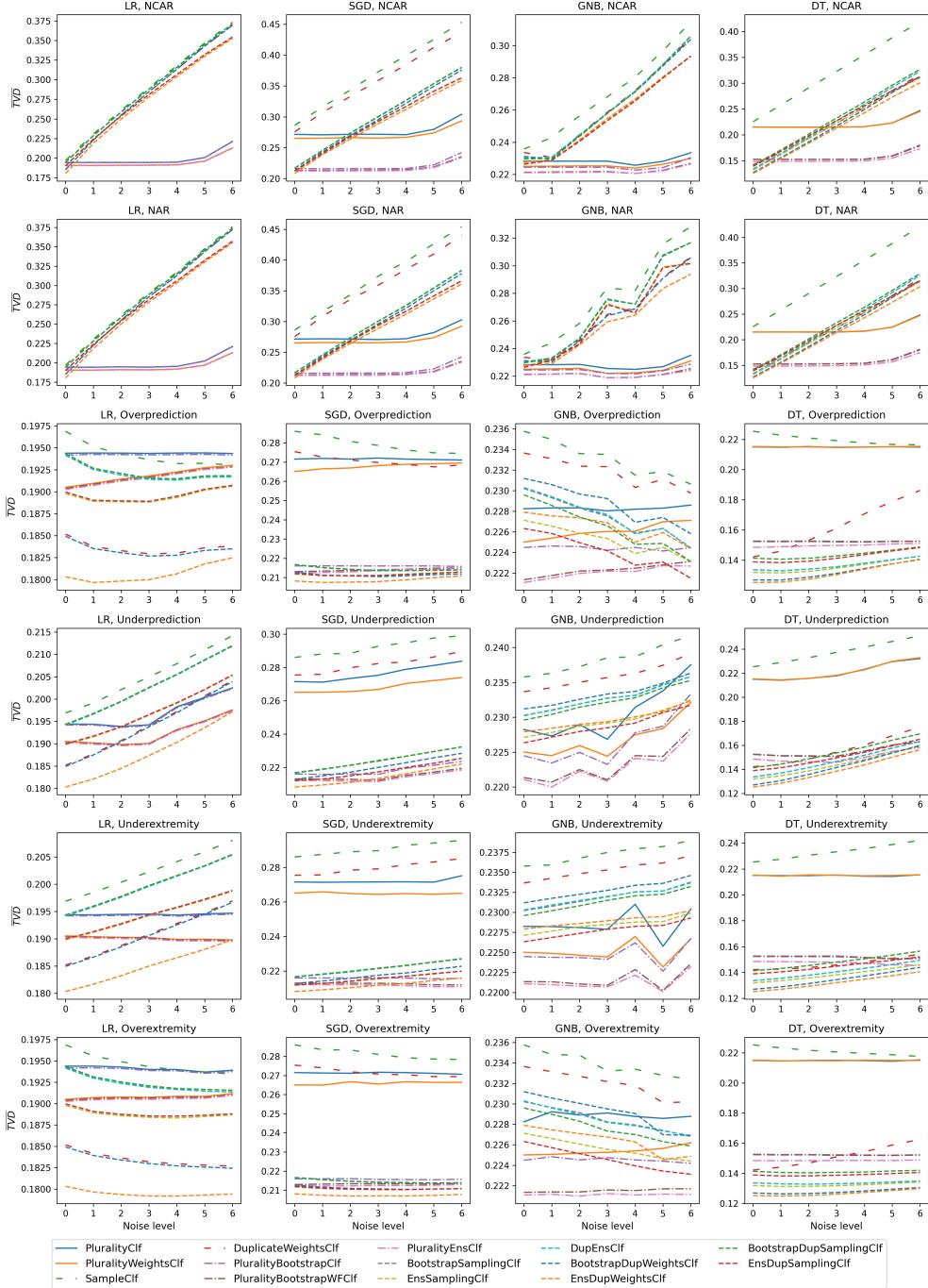


Fig. OS.8: Test data: \overline{TVD} on the partial ground truth. Uncertainty level: low.

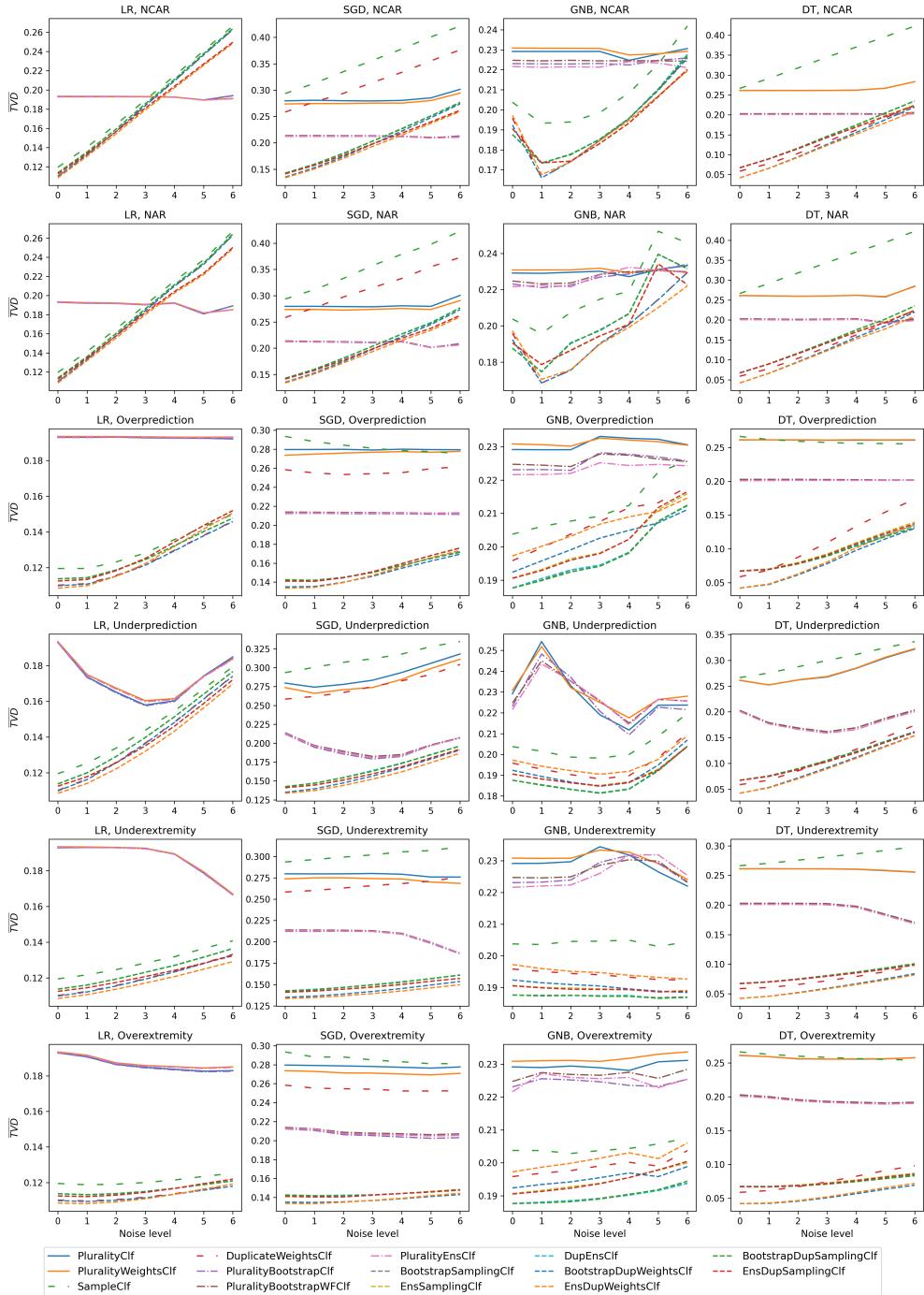


Fig. OS.9: Test data: \overline{TVD} on the partial ground truth. Uncertainty level: high.

3 Question 2: Relative Effect of Noise

Shared description for all Figures in Online Supplementary Section 3: The effect of six different noise types on method performance with four base classifiers, measured by the change in AUC or \overline{TVD} relative to the noiseless baseline (level 0), on the ground truth test data. Noise levels range from level 0 (noiseless) to level 6 (noise strength 0.3). LR, SGD, GNB and DT were used as base classifiers. RF served as the ground truth model, with the soft labels generated at either the low or high uncertainty level.

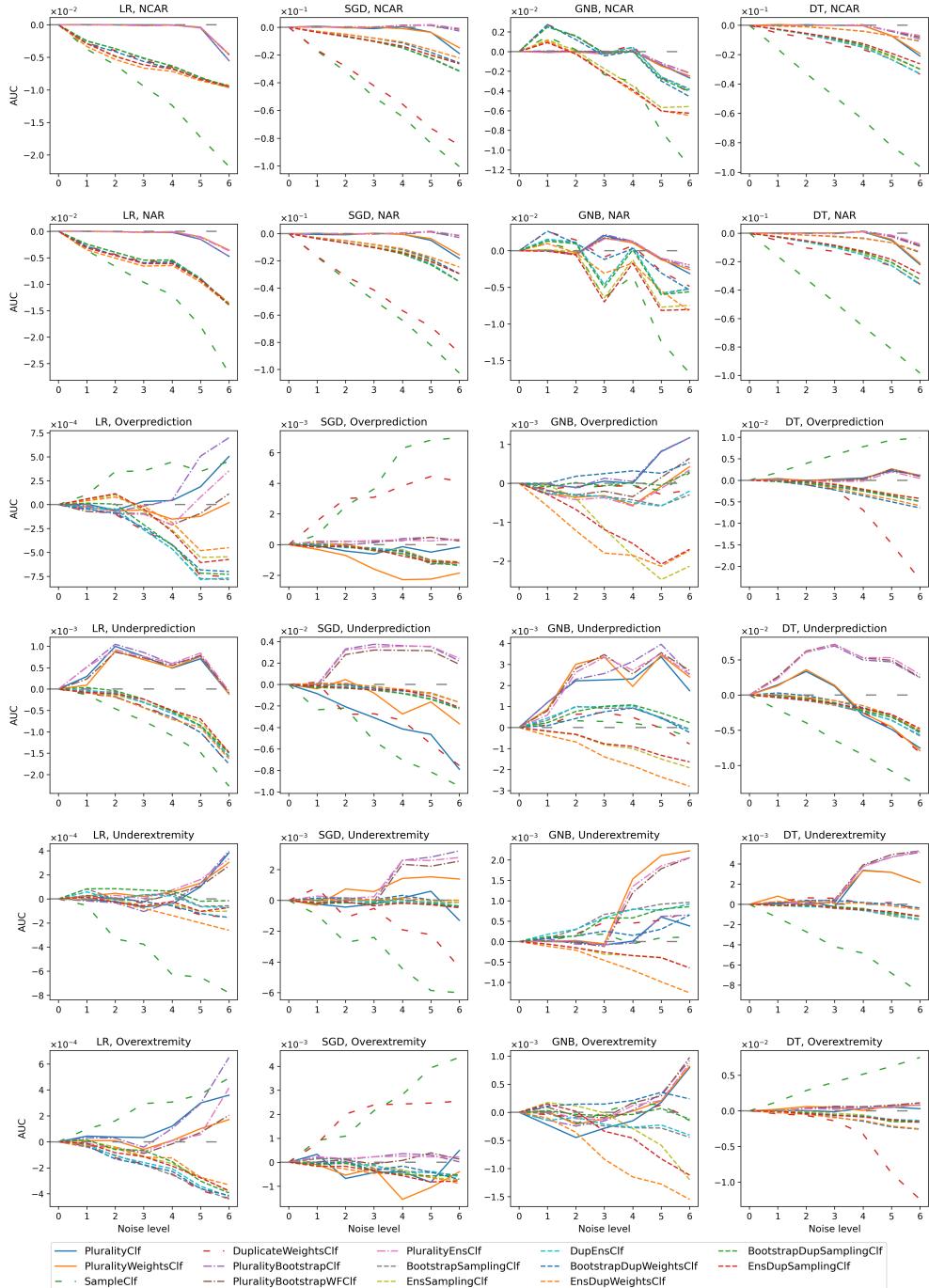


Fig. OS.10: Test data: AUC on the ground truth. Uncertainty level: low.

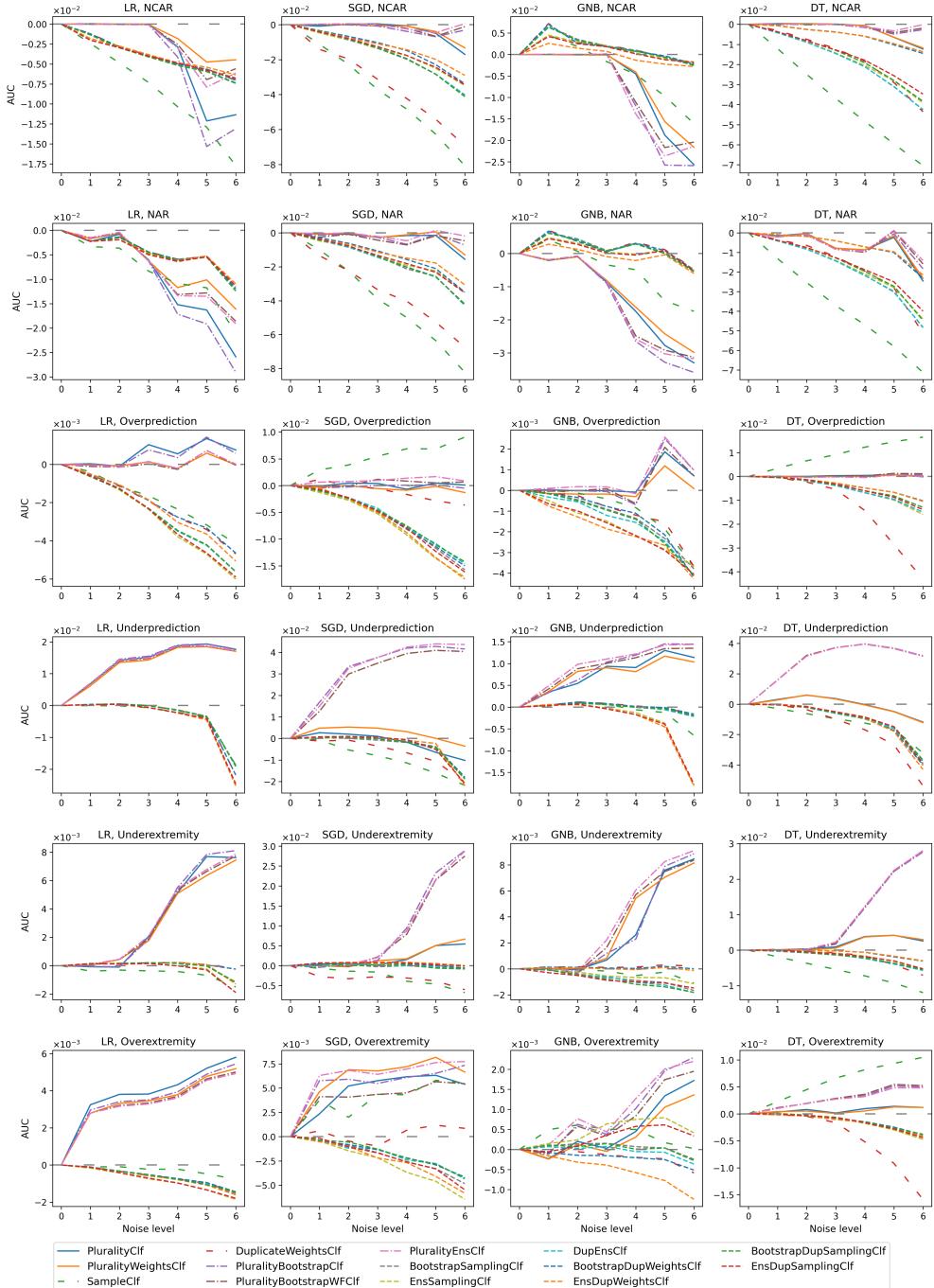


Fig. OS.11: Test data: AUC on the ground truth. Uncertainty level: high.

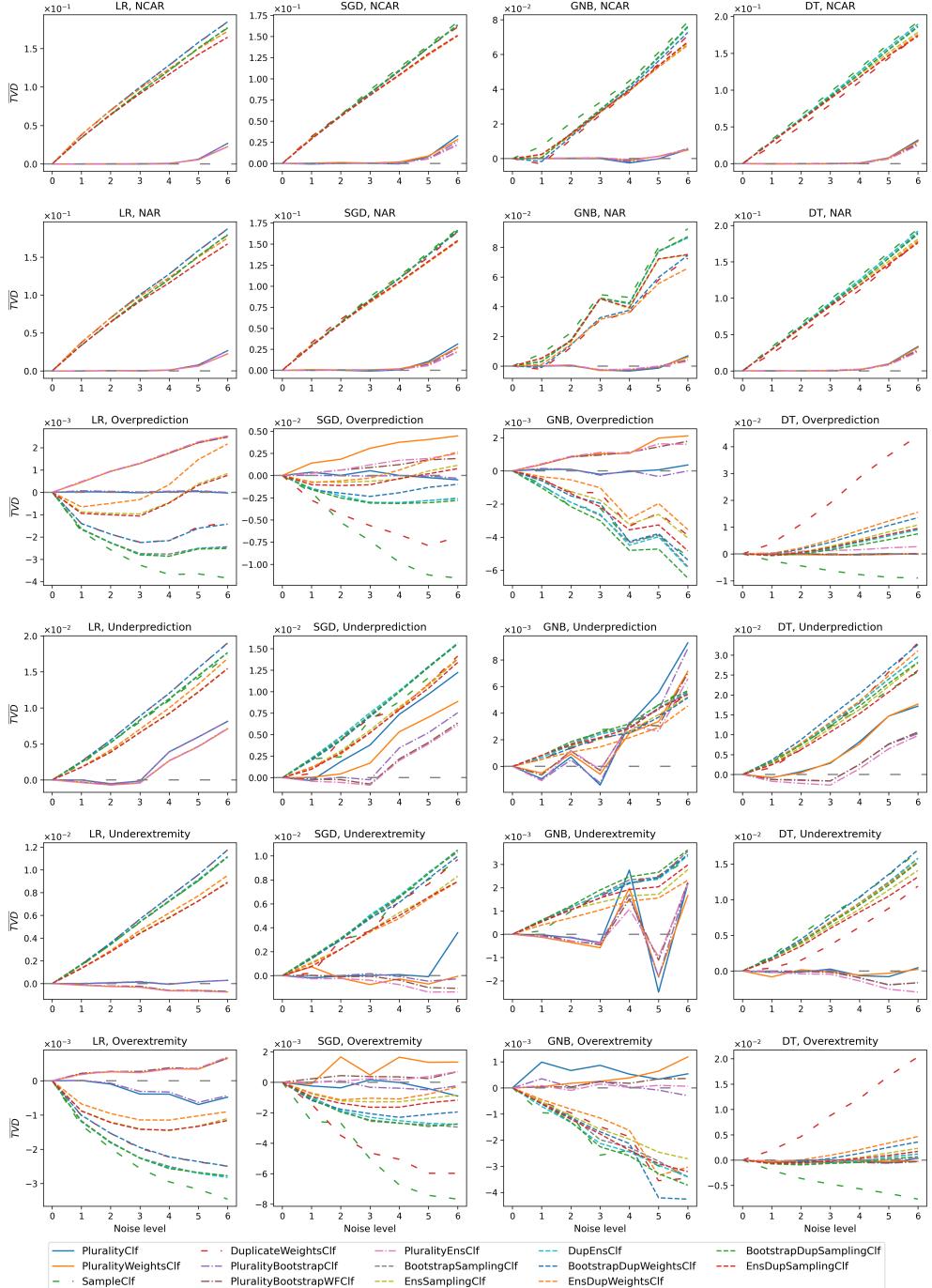


Fig. OS.12: Test data: \overline{TVD} on the partial ground truth. Uncertainty level: low.

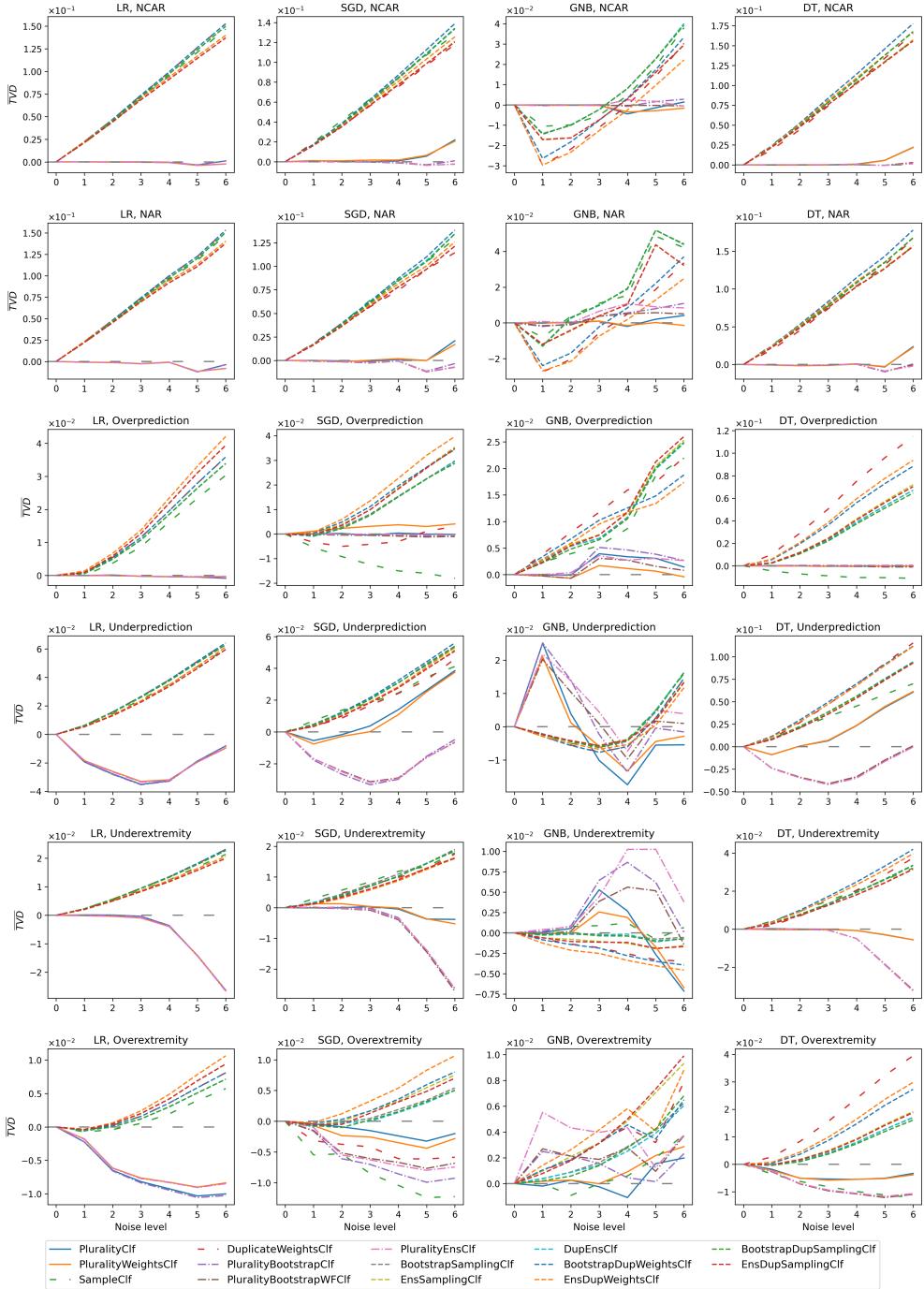


Fig. OS.13: Test data: $\overline{TV\!D}$ on the partial ground truth. Uncertainty level: high.

4 Question 2: Statistical Analyses

Shared description for all Tables in Online Supplementary Section 4: The ranks and p -values from statistical tests comparing the performance of different methods across multiple datasets. The ranks are obtained using the Friedman Aligned-Ranks test. Lower ranks indicate better performance for \overline{TVD} while higher ranks indicate better performance for AUC. Adjusted p -values from the Finner test are provided to assess the statistical significance of the differences between methods. Values below 0.05 are marked in bold, indicating statistically significant differences.

Table OS.1: Statistical analysis for low base uncertainty. Noise type: Noiseless.

	<i>p</i> -value	Rank				Adjusted <i>p</i> -value: SoftEns			
		Friedman	HardSingle	SoftSingle	HardEns	SoftEns	vs HardSingle	vs SoftSingle	vs HardEns
\overline{TVD}	LR	4.77e-01	41.94	32.82	31.18	32.06	3.75e-01	9.67e-01	9.67e-01
	SGD	5.03e-08	50.29	52.00	16.88	18.82	5.22e-06	3.00e-06	7.75e-01
	GNB	1.74e-04	32.88	48.47	15.76	40.88	3.35e-01	3.35e-01	6.38e-04
	DT	8.01e-07	58.62	29.71	31.32	18.35	8.72e-09	9.42e-02	8.26e-02
AUC	LR	4.53e-07	16.24	49.21	21.82	50.74	1.09e-06	8.22e-01	3.03e-05
	SGD	4.32e-08	18.65	16.97	49.71	52.68	7.86e-07	4.22e-07	6.61e-01
	GNB	4.29e-02	22.79	32.32	42.26	40.62	2.56e-02	3.13e-01	8.08e-01
	DT	2.60e-08	9.50	31.85	40.91	55.74	2.79e-11	6.44e-04	2.88e-02

Table OS.2: Statistical analysis for low base uncertainty. Noise type: NCAR. $\beta = 0.3$

	<i>p</i> -value	Rank				Adjusted <i>p</i> -value: SoftEns			
		Friedman	HardSingle	SoftSingle	HardEns	SoftEns	vs HardSingle	vs SoftSingle	vs HardEns
\overline{TVD}	LR	2.33e-08	18.88	51.53	16.12	51.47	2.32e-06	9.93e-01	5.59e-07
	SGD	1.92e-09	26.76	58.35	9.82	43.06	2.43e-02	2.43e-02	2.87e-06
	GNB	7.43e-07	21.24	50.06	17.18	49.53	4.53e-05	9.38e-01	5.52e-06
	DT	1.00e-07	28.65	50.12	10.53	48.71	4.65e-03	8.35e-01	5.44e-08
AUC	LR	6.51e-01	34.91	31.56	40.09	31.44	7.55e-01	9.86e-01	4.92e-01
	SGD	5.20e-09	23.82	11.41	54.59	48.18	4.95e-04	1.78e-07	3.44e-01
	GNB	1.25e-02	36.82	24.00	48.00	29.18	3.63e-01	4.45e-01	1.65e-02
	DT	2.50e-08	10.47	27.74	44.88	54.91	1.70e-10	9.23e-05	1.39e-01

Table OS.3: Statistical analysis for low base uncertainty. Noise type: NAR. $\beta = 0.3$

	<i>p</i> -value	Rank				Adjusted <i>p</i> -value: SoftEns			
		Friedman	HardSingle	SoftSingle	HardEns	SoftEns	vs HardSingle	vs SoftSingle	vs HardEns
\overline{TVD}	LR	2.32e-08	18.94	51.53	16.06	51.47	2.43e-06	9.93e-01	5.33e-07
	SGD	2.29e-09	26.35	58.35	10.18	43.12	2.01e-02	2.47e-02	3.58e-06
	GNB	5.36e-07	21.35	50.00	16.76	49.88	3.89e-05	9.86e-01	3.14e-06
	DT	1.18e-07	28.65	50.06	10.65	48.65	4.78e-03	8.35e-01	6.33e-08
AUC	LR	2.56e-02	40.18	26.35	45.18	26.29	6.04e-02	9.93e-01	1.60e-02
	SGD	4.28e-09	23.94	11.24	55.00	47.82	6.44e-04	2.06e-07	2.90e-01
	GNB	3.37e-02	35.41	26.53	47.24	28.82	4.53e-01	7.35e-01	1.98e-02
	DT	3.10e-08	10.47	27.85	45.35	54.32	3.02e-10	1.43e-04	1.86e-01

Table OS.4: Statistical analysis for high base uncertainty. Noise type: Overprediction. $\beta = 0.3$

	<i>p</i> -value	Rank				Adjusted <i>p</i> -value: SoftEns		
		Friedman	HardSingle	SoftSingle	HardEns	SoftEns	vs HardSingle	vs SoftSingle
<i>TVD</i>	LR	1.32e-03	49.47	25.12	40.24	23.18	3.17e-04	7.75e-01
	SGD	3.32e-08	52.18	50.53	17.76	17.53	9.75e-07	1.71e-06
	GNB	1.11e-03	43.71	45.41	19.35	29.53	5.65e-02	5.65e-02
	DT	1.43e-07	55.71	44.47	21.06	16.76	2.81e-08	6.61e-05
AUC	LR	1.30e-07	14.82	48.03	22.35	52.79	6.49e-08	4.82e-01
	SGD	4.01e-08	16.53	19.00	50.29	52.18	4.42e-07	1.50e-06
	GNB	8.34e-05	17.56	27.21	44.85	48.38	1.65e-05	2.69e-03
	DT	2.87e-08	10.85	27.15	45.38	54.62	3.30e-10	7.67e-05

Table OS.5: Statistical analysis for high base uncertainty. Noise type: Underprediction. $\beta = 0.3$

	<i>p</i> -value	Rank				Adjusted <i>p</i> -value: SoftEns		
		Friedman	HardSingle	SoftSingle	HardEns	SoftEns	vs HardSingle	vs SoftSingle
<i>TVD</i>	LR	3.04e-02	34.41	41.59	21.18	40.82	4.69e-01	9.10e-01
	SGD	2.35e-08	50.71	52.12	15.00	20.18	1.01e-05	7.45e-06
	GNB	1.00e-02	38.68	43.59	19.50	36.24	7.19e-01	3.87e-01
	DT	1.78e-07	59.41	35.00	21.82	21.76	8.53e-08	7.55e-02
AUC	LR	8.38e-03	22.68	42.79	27.97	44.56	3.76e-03	7.95e-01
	SGD	2.29e-08	17.35	17.65	51.03	51.97	9.97e-07	9.97e-07
	GNB	1.97e-04	19.68	26.26	48.97	43.09	1.67e-03	1.96e-02
	DT	2.49e-08	9.91	29.62	44.74	53.74	3.11e-10	5.65e-04

Table OS.6: Statistical analysis for high base uncertainty. Noise type: Underextremity. $\beta = 0.3$

	<i>p</i> -value	Rank				Adjusted <i>p</i> -value: SoftEns		
		Friedman	HardSingle	SoftSingle	HardEns	SoftEns	vs HardSingle	vs SoftSingle
<i>TVD</i>	LR	4.25e-02	33.88	41.41	21.94	40.76	4.27e-01	9.24e-01
	SGD	1.54e-08	48.88	54.00	14.29	20.82	5.28e-05	3.00e-06
	GNB	4.97e-05	29.59	50.24	16.06	42.12	9.55e-02	2.31e-01
	DT	6.39e-07	58.82	32.47	23.06	23.65	6.43e-07	2.75e-01
AUC	LR	5.00e-03	23.50	43.62	25.91	44.97	4.63e-03	8.42e-01
	SGD	3.22e-08	18.41	16.88	50.65	52.06	1.05e-06	6.43e-07
	GNB	1.67e-02	22.85	30.09	45.09	39.97	3.44e-02	2.10e-01
	DT	3.04e-08	9.85	30.68	43.21	54.26	1.75e-10	7.58e-04

Table OS.7: Statistical analysis for high base uncertainty. Noise type: Overextremity. $\beta = 0.3$

	p-value	Rank				Adjusted p-value: SoftEns			
		Friedman	HardSingle	SoftSingle	HardEns	SoftEns	vs HardSingle	vs SoftSingle	vs HardEns
\overline{TVD}	LR	3.77e-02	47.53	28.71	34.41	27.35	8.77e-03	8.42e-01	4.12e-01
	SGD	3.37e-08	51.53	51.18	17.35	17.94	2.20e-06	2.20e-06	9.31e-01
	GNB	1.33e-04	42.35	46.47	14.82	34.35	2.38e-01	1.09e-01	1.19e-02
	DT	1.59e-07	58.88	36.35	26.82	15.94	7.29e-10	3.92e-03	1.09e-01
AUC	LR	9.29e-08	14.35	48.62	22.53	52.50	5.58e-08	5.67e-01	1.49e-05
	SGD	4.28e-08	17.76	17.79	50.06	52.38	9.97e-07	9.97e-07	7.32e-01
	GNB	1.48e-03	19.15	30.09	44.09	44.68	5.01e-04	4.69e-02	9.31e-01
	DT	2.53e-08	10.15	29.03	43.50	55.32	8.16e-11	1.59e-04	8.13e-02

5 Real World Data Tables

Table OS.8: Method performance on the UrinCheck dataset, measured by AUC. The labels used for the test set are the most probable labels from the soft-labelled data. The numbers are visualised in Figure 6 of the accompanying work. The results have been multiplied by 10^2 to enhance readability.

Method	LR						SGD					
	5	10	20	40	60	80	5	10	20	40	60	80
HardSingle												
PluralityClf	86.40	88.04	89.08	89.67	89.87	89.96	74.20	74.65	74.86	75.32	76.02	76.74
PluralityWeightsClf	86.62	88.20	89.15	89.70	89.91	90.00	74.56	75.15	75.59	75.67	76.49	76.94
SoftSingle												
SampleClf	77.86	82.49	86.09	88.21	88.85	89.24	63.88	64.93	65.05	65.37	65.12	66.85
DuplicateWeightsClf	86.60	88.49	89.38	89.83	90.02	90.11	65.92	65.75	65.16	66.39	67.12	68.56
HardEns												
PluralityEnsClf	87.65	88.66	89.37	89.80	89.97	90.04	85.90	87.43	88.83	89.62	89.80	90.00
PluralityBootstrapClf	87.30	88.51	89.27	89.74	89.93	90.00	85.58	87.21	88.75	89.56	89.79	89.96
PluralityBootstrapWFClf	87.57	88.62	89.35	89.79	89.96	90.03	85.76	87.32	88.83	89.60	89.81	89.98
DupEnsClf	87.42	88.60	89.36	89.84	90.00	90.08	84.71	86.44	87.38	87.99	88.26	88.31
SoftEns												
BootstrapSamplingClf	87.33	88.69	89.39	89.82	90.00	90.07	84.44	86.49	87.35	87.96	88.18	88.32
EnsSamplingClf	87.49	88.62	89.40	89.82	90.01	90.08	84.59	86.66	87.56	88.09	88.37	88.71
BootstrapDupSamplingClf	87.51	88.68	89.42	89.83	89.98	90.08	84.21	86.37	87.38	88.09	88.32	88.38
EnsDupSamplingClf	87.45	88.70	89.42	89.84	89.99	90.07	84.76	86.52	87.48	88.12	88.40	88.59
BootstrapDupWeightsClf	87.10	88.59	89.39	89.84	90.02	90.10	84.86	86.64	87.53	88.11	88.43	88.85
EnsDupWeightsClf	87.14	88.60	89.40	89.84	90.01	90.09	85.03	86.71	87.78	88.41	88.58	88.93
Method	GNB						DT					
	5	10	20	40	60	80	5	10	20	40	60	80
HardSingle												
PluralityClf	87.88	89.61	90.09	90.23	90.26	90.36	71.78	72.95	74.46	75.69	76.55	77.34
PluralityWeightsClf	87.93	89.65	90.11	90.24	90.26	90.37	71.50	73.08	74.43	75.67	76.55	77.22
SoftSingle												
SampleClf	82.36	87.10	89.07	89.73	89.86	90.00	63.16	63.90	64.46	65.19	65.71	66.14
DuplicateWeightsClf	88.54	89.57	89.88	90.00	90.02	90.13	75.08	76.76	78.31	79.88	80.80	81.70
HardEns												
PluralityEnsClf	86.95	89.46	90.15	90.32	90.32	90.42	86.80	87.89	88.73	89.45	89.90	90.31
PluralityBootstrapClf	86.94	89.39	90.13	90.30	90.31	90.40	86.64	87.82	88.59	89.38	89.79	90.29
PluralityBootstrapWFClf	87.08	89.49	90.15	90.32	90.32	90.41	86.67	87.80	88.63	89.36	89.81	90.30
DupEnsClf	87.62	89.45	89.87	90.01	90.03	90.13	85.05	86.25	87.07	87.77	88.11	88.67
SoftEns												
BootstrapSamplingClf	87.60	89.46	89.88	90.01	90.02	90.12	84.92	86.23	87.05	87.79	88.08	88.62
EnsSamplingClf	87.69	89.47	89.87	90.00	90.02	90.12	85.17	86.28	87.17	87.80	88.29	88.73
BootstrapDupSamplingClf	87.35	89.38	89.87	90.02	90.03	90.13	85.49	86.68	87.61	88.25	88.72	89.02
EnsDupSamplingClf	87.26	89.38	89.86	90.01	90.02	90.12	85.53	86.77	87.66	88.41	88.77	89.18
BootstrapDupWeightsClf	88.78	89.59	89.89	90.00	90.02	90.13	87.00	88.01	88.77	89.44	89.91	90.38
EnsDupWeightsClf	88.84	89.61	89.88	89.99	90.01	90.12	87.04	88.04	88.80	89.47	89.92	90.45

Table OS.9: Method performance on the UrinCheck dataset, measured by AUC. The labels used for the test set are the most probable labels from the soft-labelled data. The numbers are visualised in Figure F3 in Appendix F of the accompanying work. The results have been multiplied by 10^2 to enhance readability.

Method	LR						SGD					
	5	10	20	40	60	80	5	10	20	40	60	80
HardSingle												
PluralityClf	70.07	71.02	71.68	71.95	72.10	72.30	63.20	63.56	63.59	63.92	64.17	64.62
PluralityWeightsClf	70.22	71.23	71.69	72.16	72.15	72.11	63.46	63.83	63.93	64.06	64.45	64.72
SoftSingle												
SampleClf	65.36	67.99	69.89	71.36	71.50	72.03	57.67	58.24	58.10	58.60	58.34	59.27
DuplicateWeightsClf	70.17	71.48	71.89	72.23	72.39	72.30	58.63	58.69	58.36	58.95	59.35	60.20
HardEns												
PluralityEnsClf	70.72	71.42	71.79	72.09	72.00	72.28	69.61	70.39	71.22	71.60	71.86	71.83
PluralityBootstrapClf	70.44	71.22	71.77	71.96	72.19	72.10	69.34	70.27	71.16	71.61	71.87	71.81
PluralityBootstrapWFClf	70.73	71.32	71.76	72.05	72.10	72.04	69.42	70.42	71.34	71.64	71.78	71.77
DupEnsClf	70.65	71.41	71.90	72.15	72.22	72.43	69.05	70.05	70.55	70.79	71.00	71.06
SoftEns												
BootstrapSamplingClf	70.61	71.52	71.92	72.14	72.25	72.30	68.90	70.09	70.59	71.01	70.80	71.25
EnsSamplingClf	70.77	71.37	72.00	72.30	72.53	72.49	69.00	70.14	70.63	71.03	71.16	70.99
BootstrapDupSamplingClf	70.69	71.51	71.91	72.23	72.30	72.39	68.81	70.03	70.53	70.94	71.04	70.84
EnsDupSamplingClf	70.75	71.48	71.92	72.24	72.28	72.31	69.13	70.00	70.61	70.84	71.09	71.16
BootstrapDupWeightsClf	70.53	71.42	71.97	72.18	72.40	72.45	69.10	70.10	70.65	70.84	71.22	71.52
EnsDupWeightsClf	70.58	71.49	72.02	72.19	72.27	72.45	69.34	70.28	70.80	71.15	71.29	71.36
GNB												
Method	5	10	20	40	60	80	5	10	20	40	60	80
	70.96	72.02	72.17	72.37	72.38	72.66	61.84	62.43	63.28	63.93	64.41	64.81
PluralityClf	71.07	71.98	72.31	72.44	72.35	72.58	61.71	62.58	63.21	63.96	64.29	64.72
SoftSingle												
SampleClf	67.91	70.65	71.71	72.16	72.21	72.06	57.25	57.70	57.87	58.44	58.72	58.95
DuplicateWeightsClf	71.39	72.03	72.14	72.31	72.34	72.45	63.98	64.94	65.91	66.69	67.31	67.70
HardEns												
PluralityEnsClf	70.31	71.95	72.26	72.40	72.19	72.36	70.08	70.82	71.13	71.53	71.82	71.92
PluralityBootstrapClf	70.46	71.81	72.40	72.35	72.28	72.22	70.07	70.77	71.05	71.46	71.70	71.76
PluralityBootstrapWFClf	70.59	71.89	72.41	72.34	72.47	72.25	70.08	70.68	71.04	71.62	71.76	71.94
DupEnsClf	70.75	71.80	72.21	72.28	72.32	72.31	69.33	70.11	70.49	70.79	71.14	71.47
SoftEns												
BootstrapSamplingClf	70.74	71.76	72.19	72.22	72.29	72.19	69.21	70.07	70.46	70.81	70.90	71.46
EnsSamplingClf	70.78	71.87	72.19	72.45	72.36	72.39	69.37	69.97	70.56	70.97	71.17	71.32
BootstrapDupSamplingClf	70.51	71.83	72.21	72.25	72.42	72.29	69.58	70.24	70.84	71.16	71.57	71.63
EnsDupSamplingClf	70.51	71.88	72.15	72.40	72.29	72.31	69.45	70.21	70.82	71.22	71.44	71.80
BootstrapDupWeightsClf	71.45	72.04	72.28	72.46	72.45	72.49	70.43	71.03	71.66	71.90	72.35	72.44
EnsDupWeightsClf	71.46	71.98	72.26	72.41	72.40	72.52	70.44	71.12	71.61	72.01	72.20	72.35

Table OS.10: Method performance on the UrinCheck dataset, measured by \overline{TVD} . The labels used for the test set are the most probable labels from the soft-labelled data. The numbers are visualised in Figure 7 of the accompanying work. The results have been multiplied by 10^2 to enhance readability.

Method	LR						SGD					
	5	10	20	40	60	80	5	10	20	40	60	80
HardSingle												
PluralityClf	20.64	20.33	19.52	18.98	18.77	18.71	33.92	33.37	33.25	33.05	32.55	32.11
PluralityWeightsClf	20.62	20.43	19.73	19.26	19.08	19.03	33.68	33.13	32.91	32.79	32.30	31.96
SoftSingle												
SampleClf	20.46	17.88	16.08	15.03	14.66	14.43	40.62	39.87	39.58	39.12	38.85	37.35
DuplicateWeightsClf	15.58	14.70	14.25	14.00	13.88	13.83	39.35	39.23	39.34	37.83	36.25	34.30
HardEns												
PluralityEnsClf	19.61	20.02	19.63	19.26	19.09	19.05	23.17	22.71	21.36	20.50	20.24	20.10
PluralityBootstrapClf	19.64	19.93	19.44	18.99	18.79	18.72	23.27	22.62	21.21	20.24	19.96	19.77
PluralityBootstrapWFClf	19.62	20.02	19.63	19.26	19.08	19.04	23.27	22.71	21.40	20.47	20.21	20.03
DupEnsClf	15.76	14.90	14.39	14.07	13.94	13.87	17.48	16.64	16.18	15.95	15.76	15.79
SoftEns												
BootstrapSamplingClf	15.78	14.90	14.38	14.08	13.94	13.88	17.60	16.66	16.21	15.93	15.79	15.74
EnsSamplingClf	15.62	14.83	14.26	13.96	13.81	13.75	17.52	16.59	16.09	15.89	15.70	15.57
BootstrapDupSamplingClf	15.89	14.98	14.40	14.09	13.95	13.88	17.61	16.70	16.22	15.86	15.78	15.73
EnsDupSamplingClf	15.79	14.86	14.29	13.96	13.82	13.75	17.50	16.63	16.18	15.85	15.73	15.61
BootstrapDupWeightsClf	15.50	14.71	14.26	14.00	13.89	13.83	17.87	16.73	16.22	15.89	15.70	15.44
EnsDupWeightsClf	15.41	14.61	14.15	13.89	13.76	13.71	17.87	16.76	16.14	15.77	15.58	15.39
GNB												
Method	5	10	20	40	60	80	5	10	20	40	60	80
	29.54	27.09	25.48	25.01	24.93	24.86	34.97	34.23	33.44	32.87	32.47	32.15
HardSingle												
PluralityClf	29.54	27.21	25.65	25.20	25.12	25.06	35.10	34.14	33.47	32.89	32.48	32.17
SoftSingle												
SampleClf	27.52	22.36	20.30	19.95	19.77	19.67	40.82	40.30	39.88	39.33	38.85	38.35
DuplicateWeightsClf	20.65	20.11	19.80	19.66	19.60	19.57	18.29	17.54	16.89	16.28	15.93	15.63
HardEns												
PluralityEnsClf	23.83	25.06	24.70	24.84	24.91	24.91	20.47	20.77	20.86	20.98	21.09	21.17
PluralityBootstrapClf	23.99	24.96	24.60	24.67	24.73	24.72	20.42	20.77	20.87	20.97	21.10	21.15
PluralityBootstrapWFClf	23.84	24.99	24.73	24.86	24.92	24.91	20.47	20.76	20.85	20.96	21.10	21.14
DupEnsClf	17.59	18.29	18.99	19.30	19.34	19.38	16.94	16.24	15.77	15.39	15.19	14.96
SoftEns												
BootstrapSamplingClf	17.68	18.31	19.00	19.30	19.35	19.38	16.98	16.23	15.79	15.41	15.20	14.97
EnsSamplingClf	17.72	18.71	19.47	19.78	19.82	19.85	16.83	16.17	15.71	15.36	15.11	14.92
BootstrapDupSamplingClf	16.92	17.72	18.73	19.19	19.28	19.33	16.87	16.11	15.59	15.19	14.93	14.76
EnsDupSamplingClf	17.13	17.97	19.19	19.66	19.75	19.80	16.78	15.99	15.50	15.09	14.86	14.64
BootstrapDupWeightsClf	19.41	19.65	19.60	19.57	19.53	19.52	15.74	14.96	14.45	14.05	13.80	13.58
EnsDupWeightsClf	19.79	20.06	20.07	20.04	19.99	19.97	15.64	14.87	14.38	13.98	13.74	13.52
DT												