Supplementary Material 1 to the article:

Diversity Forests: Using Split Sampling to Allow for Complex Split Procedures in Random Forest

Roman Hornung*,[1]

[1] Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany

*Contact: hornung@ibe.med.uni-muenchen.de

A Overview of the Data Sets used in the Analyses

data.id	label	n	p	prop. categ.	prop. min. class
31	credit-g	1000	20	0.650	0.300
37	diabetes	768	8	0.000	0.349
40	sonar	208	60	0.000	0.466
43	haberman	306	3	0.333	0.265
44	spambase	4601	57	0.000	0.394
50	tic-tac-toe	958	9	1.000	0.347
53	heart-statlog	270	13	0.000	0.444
59	ionosphere	351	34	0.000	0.359
164	molecular-biology_promoters	106	57	1.000	0.500
292	Australian	690	14	0.000	0.445
311	oil_spill	937	49	0.000	0.044
312	scene	2407	299	0.017	0.179
316	yeast_ml8	2417	116	0.112	0.014
333	monks-problems-1	556	6	1.000	0.500
334	monks-problems-2	601	6	1.000	0.343
335	monks-problems-3	554	6	1.000	0.480
336	SPECT	267	22	1.000	0.206
337	SPECTF	349	44	0.000	0.272
346	aids	50	4	0.500	0.500
444	analcatdata_boxing2	132	3	1.000	0.462
446	prnn_crabs	200	7	0.143	0.500
448	analcatdata_boxing1	120	3	1.000	0.350
450	analcatdata_lawsuit	264	4	0.250	0.072
459	analcatdata_asbestos	83	3	0.667	0.446
467	analcatdata_japansolvent	52	8	0.000	0.481
472	lupus	87	3	0.000	0.402
476	analcatdata_bankruptcy	50	5	0.000	0.500
479	analcatdata_cyyoung9302	92	9	0.333	0.207
682	sleuth_ex2016	87	10	0.100	0.414
683	sleuth_ex2015	60	7	0.000	0.500
713	vineyard	52	3	0.000	0.462
714	fruitfly	125	4	0.500	0.392
717	rmftsa_ladata	508	10	0.000	0.437
719	veteran	137	7	0.571	0.314
720	abalone	4177	8	0.125	0.498
721	pwLinear	200	10	0.000	0.485
724	analcatdata_vineyard	468	3	0.333	0.444
725	bank8FM	8192	8	0.000	0.404
728	analcatdata_supreme	4052	7	0.000	0.240
729	visualizing_slope	44	3	0.000	0.386
731	baskball	96	4	0.000	0.490
733	machine_cpu	209	6	0.000	0.268
735	cpu_small	8192	12	0.000	0.302

Table S1: Overview of data sets – I. The following information is provided: ‘data.id’: OpenML ID of the data set, ‘label’: data set label, ‘n’: sample size, ‘p’: number of features, ‘prop. categ.’: proportion of categorial features, ‘prop. min. class’: proportion of observations in the smaller class of the target variable.

data.id	label	n	p	prop. categ.	prop. min. class
736	visualizing_environmental	111	3	0.000	0.477
737	space_ga	3107	6	0.000	0.496
741	rmftsa_sleepdata	1024	2	0.500	0.497
745	auto_price	159	15	0.067	0.340
747	servo	167	4	1.000	0.228
748	analcatdata_wildcat	163	5	0.400	0.288
750	pm10	500	7	0.000	0.492
753	wisconsin	194	32	0.000	0.464
755	sleuth_ex1605	62	5	0.000	0.500
758	analcatdata_election2000	67	14	0.000	0.269
759	analcatdata_olympic2000	66	11	0.000	0.500
761	cpu_act	8192	21	0.000	0.302
764	analcatdata_apnea3	450	3	0.667	0.122
765	analcatdata_apnea2	475	3	0.667	0.135
767	analcatdata_apnea1	475	3	0.667	0.128
770	strikes	625	6	0.000	0.496
771	analcatdata_michiganacc	108	4	0.500	0.444
772	quake	2178	3	0.000	0.445
774	disclosure_x_bias	662	3	0.000	0.479
777	sleuth_ex1714	47	7	0.000	0.426
778	bodyfat	252	14	0.000	0.492
780	rabe_265	51	6	0.000	0.412
782	rabe_266	120	2	0.000	0.475
784	newton_hema	140	3	0.333	0.500
787	witmer_census_1980	50	4	0.000	0.480
788	triazines	186	60	0.000	0.414
790	elusage	55	2	0.500	0.436
791	diabetes_numeric	43	2	0.000	0.395
795	disclosure_x_tampered	662	3	0.000	0.494
800	pyrim	74	27	0.000	0.419
801	chscase_funds	185	2	0.000	0.470
803	delta_ailerons	7129	5	0.000	0.469
804	hutsof99_logis	70	7	0.571	0.486
807	kin8nm	8192	8	0.000	0.491
811	rmftsa_ctoarrivals	264	2	0.500	0.383
814	chscase_vine2	468	2	0.000	0.453
815	chscase_vine1	52	9	0.000	0.462
816	puma8NH	8192	8	0.000	0.498
817	diggle_table_a1	48	4	0.000	0.479
818	diggle_table_a2	310	8	0.125	0.468
819	delta_elevators	9517	6	0.000	0.497
820	chatfield_4	235	12	0.000	0.396
826	sensory	576	11	1.000	0.415
827	disclosure_x_noise	662	3	0.000	0.497
835	analcatdata_vehicle	48	4	1.000	0.438

Table S2: Overview of data sets – II. The following information is provided: ‘data.id’: OpenML ID of the data set, ‘label’: data set label, ‘n’: sample size, ‘p’: number of features, ‘prop. categ.’: proportion of categorial features, ‘prop. min. class’: proportion of observations in the smaller class of the target variable.

data.id	label	n	p	prop. categ.	prop. min. class
841	stock	950	9	0.000	0.486
847	wind	6574	14	0.000	0.467
848	schlvote	38	5	0.200	0.263
851	tecator	240	124	0.000	0.425
853	boston	506	13	0.077	0.413
857	bolts	40	7	0.000	0.350
859	analcatdata_gviolence	74	8	0.000	0.419
860	vinnie	380	2	0.000	0.487
872	boston	506	13	0.154	0.413
874	rabe_131	50	5	0.000	0.420
875	analcatdata_chlamydia	100	3	1.000	0.190
880	mu284	284	10	0.000	0.500
882	pollution	60	15	0.000	0.483
885	transplant	131	3	0.000	0.366
886	no2	500	7	0.000	0.498
887	mbagrade	61	2	0.500	0.475
890	cloud	108	7	0.143	0.296
892	sleuth_case1201	50	6	0.000	0.480
893	visualizing_hamster	73	5	0.000	0.452
894	rabe_148	66	5	0.000	0.500
895	chscase_geyser1	222	2	0.000	0.396
900	chscase_census6	400	6	0.000	0.412
902	sleuth_case2002	147	6	0.667	0.469
905	chscase_adopt	39	2	0.000	0.308
906	chscase_census5	400	7	0.000	0.482
907	chscase_census4	400	7	0.000	0.485
908	chscase_census3	400	7	0.000	0.480
909	chscase_census2	400	7	0.000	0.492
914	balloon	2001	2	0.000	0.241
915	plasma_retinol	315	13	0.231	0.422
919	rabe_166	40	2	0.000	0.475
921	analcatdata_seropositive	132	3	0.333	0.348
923	visualizing_soil	8641	4	0.250	0.450
924	humandevel	130	2	0.000	0.500
925	visualizing_galaxy	323	4	0.000	0.458
927	hutsof99_child_witness	42	16	0.000	0.405
928	rabe_97	46	4	0.250	0.457
929	rabe_176	70	4	0.000	0.500
931	disclosure_z	662	3	0.000	0.474
934	socmob	1156	5	0.800	0.221
945	kidney	76	6	0.500	0.474
946	visualizing_ethanol	88	2	0.000	0.489
947	arsenic-male-bladder	559	3	0.000	0.043
949	arsenic-female-bladder	559	3	0.000	0.143
950	arsenic-female-lung	559	3	0.000	0.034

Table S3: Overview of data sets – III. The following information is provided: 'data.id': OpenML ID of the data set, 'label': data set label, 'n': sample size, 'p': number of features, 'prop. categ.': proportion of categorial features, 'prop. min. class': proportion of observations in the smaller class of the target variable.

data.id	label	n	p	prop. categ.	prop. min. class
951	arsenic-male-lung	559	3	0.000	0.023
954	spectrometer	531	101	0.010	0.104
955	tae	151	5	0.400	0.344
958	segment	2310	19	0.000	0.143
962	mfeat-morphological	2000	6	0.000	0.100
964	pasture	36	22	0.045	0.333
965	zoo	101	16	0.938	0.406
969	iris	150	4	0.000	0.333
970	analcatdata_authorship	841	70	0.000	0.377
971	mfeat-fourier	2000	76	0.000	0.100
973	wine	178	13	0.000	0.399
974	hayes-roth	132	4	0.000	0.386
976	JapaneseVowels	9961	14	0.000	0.162
978	mfeat-factors	2000	216	0.000	0.100
980	optdigits	5620	64	0.000	0.102
983	cmc	1473	9	0.778	0.427
987	collins	500	22	0.091	0.160
988	fl2000	67	15	0.067	0.388
991	car	1728	6	1.000	0.300
994	vehicle	846	18	0.000	0.258
995	mfeat-zernike	2000	47	0.000	0.100
997	balance-scale	625	4	0.000	0.461
1005	glass	214	9	0.000	0.355
1009	white-clover	63	31	0.129	0.397
1011	ecoli	336	7	0.000	0.426
1013	analcatdata_challenger	138	2	0.500	0.065
1014	analcatdata_dmft	797	4	1.000	0.194
1015	confidence	72	3	0.000	0.167
1016	vowel	990	13	0.231	0.091
1020	mfeat-karhunen	2000	64	0.000	0.100
1021	page-blocks	5473	10	0.000	0.102
1022	mfeat-pixel	2000	240	1.000	0.100
1025	analcatdata_germangss	400	5	1.000	0.225
1043	ada_agnostic	4562	48	0.000	0.248
1045	kc1-top5	145	94	0.000	0.055
1048	jEdit_4.2_4.3	369	8	0.000	0.447
1049	pc4	1458	37	0.000	0.122
1050	pc3	1563	37	0.000	0.102
1054	mc2	161	39	0.000	0.323
1055	cm1_req	89	8	0.125	0.225
1056	mc1	9466	38	0.000	0.007
1059	ar1	121	29	0.000	0.074
1060	ar3	63	29	0.000	0.127
1061	ar4	107	29	0.000	0.187
1062	ar5	36	29	0.000	0.222

Table S4: Overview of data sets – IV. The following information is provided: 'data.id': OpenML ID of the data set, 'label': data set label, 'n': sample size, 'p': number of features, 'prop. categ.': proportion of categorial features, 'prop. min. class': proportion of observations in the smaller class of the target variable.

data.id	label	n	p	prop. categ.	prop. min. class
1063	kc2	522	21	0.000	0.205
1064	ar6	101	29	0.000	0.149
1065	kc3	458	39	0.000	0.094
1066	kc1-binary	145	94	0.000	0.414
1067	kc1	2109	21	0.000	0.155
1068	pc1	1109	21	0.000	0.069
1069	pc2	5589	36	0.000	0.004
1071	mw1	403	37	0.000	0.077
1073	jEdit_4.0_4.2	274	8	0.000	0.489
1075	datatrieve	130	8	0.000	0.085
1121	badges2	294	11	0.273	0.286
1441	KungChi3	123	39	0.000	0.130
1442	MegaWatt1	253	37	0.000	0.107
1443	PizzaCutter1	661	37	0.000	0.079
1444	PizzaCutter3	1043	37	0.000	0.122
1446	CostaMadre1	296	37	0.000	0.128
1447	CastMetal1	327	37	0.000	0.128
1451	PieChart1	705	37	0.000	0.087
1452	PieChart2	745	36	0.000	0.021
1453	PieChart3	1077	37	0.000	0.124
1454	PieChart4	1458	37	0.000	0.122
1455	acute-inflammations	120	6	0.833	0.417
1462	banknote-authentication	1372	4	0.000	0.445
1463	blogger	100	5	1.000	0.320
1464	blood-transfusion-service-center	748	4	0.000	0.238
1467	climate-model-simulation-crashes	540	20	0.000	0.085
1473	fertility	100	9	0.000	0.120
1479	hill-valley	1212	100	0.000	0.500
1480	ilpd	583	10	0.100	0.286
1487	ozone-level-8hr	2534	72	0.000	0.063
1488	parkinsons	195	22	0.000	0.246
1489	phoneme	5404	5	0.000	0.293
1490	planning-relax	182	12	0.000	0.286
1494	qsar-biodeg	1055	41	0.000	0.337
1495	qualitative-bankruptcy	250	6	1.000	0.428
1498	sa-heart	462	9	0.111	0.346
1504	steel-plates-fault	1941	33	0.000	0.347
1510	wdbc	569	30	0.000	0.373
1511	wholesale-customers	440	8	0.125	0.323
1524	vertebra-column	310	6	0.000	0.323
1547	autoUniv-au1-1000	1000	20	0.000	0.259
1570	wilt	4839	5	0.000	0.054

Table S5: Overview of data sets – V. The following information is provided: ‘data.id’: OpenML ID of the data set, ‘label’: data set label, ‘n’: sample size, ‘p’: number of features, ‘prop. categ.’: proportion of categorial features, ‘prop. min. class’: proportion of observations in the smaller class of the target variable.

B Pre-study: Cross-validated AUC Values Obtained for the Different Tuning Parameter Values and Data Sets

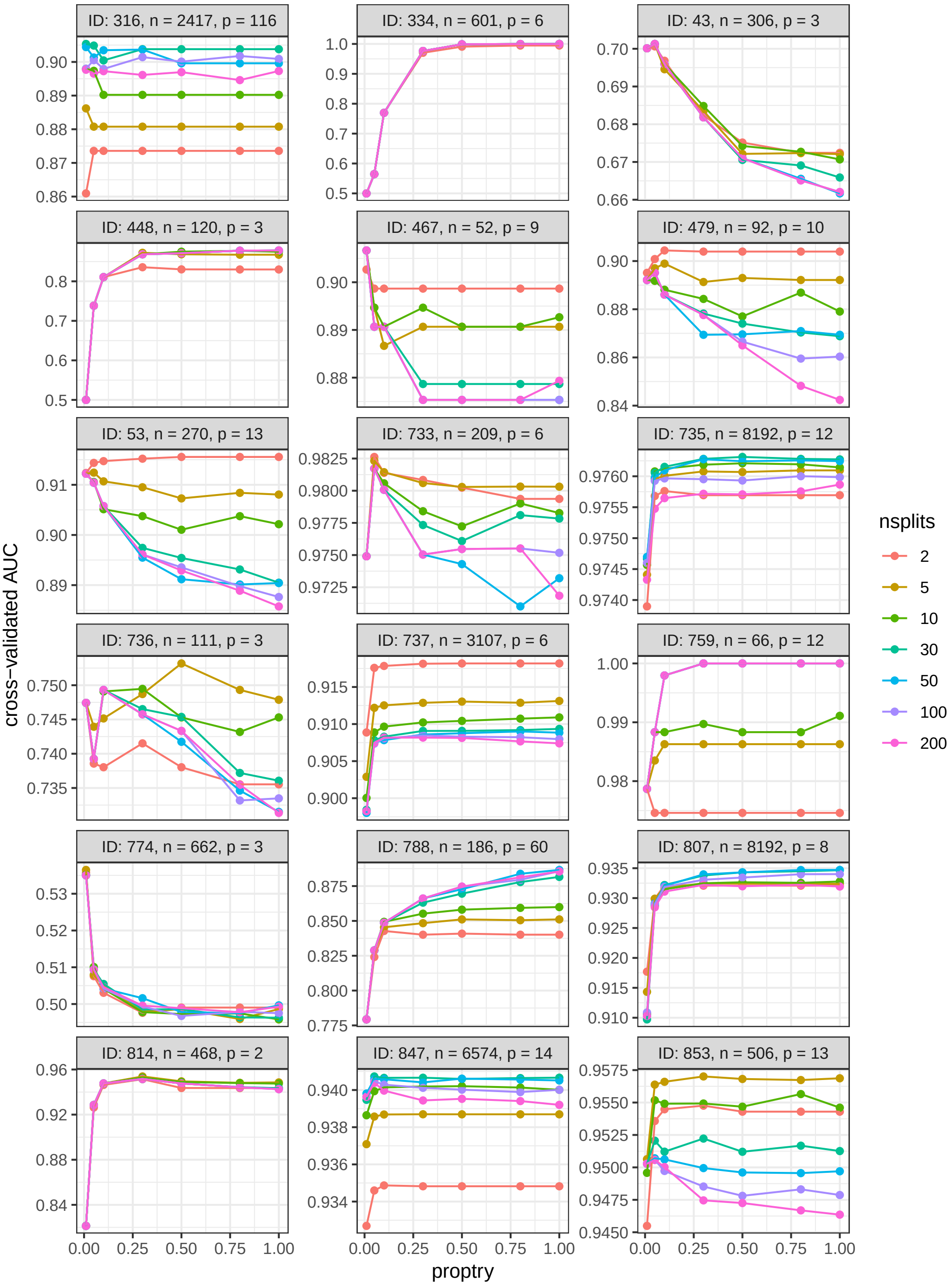

Fig. S1: Pre-study: Cross-validated AUC values obtained for the different parameter values. Each panel shows the results obtained for a particular data set. – I

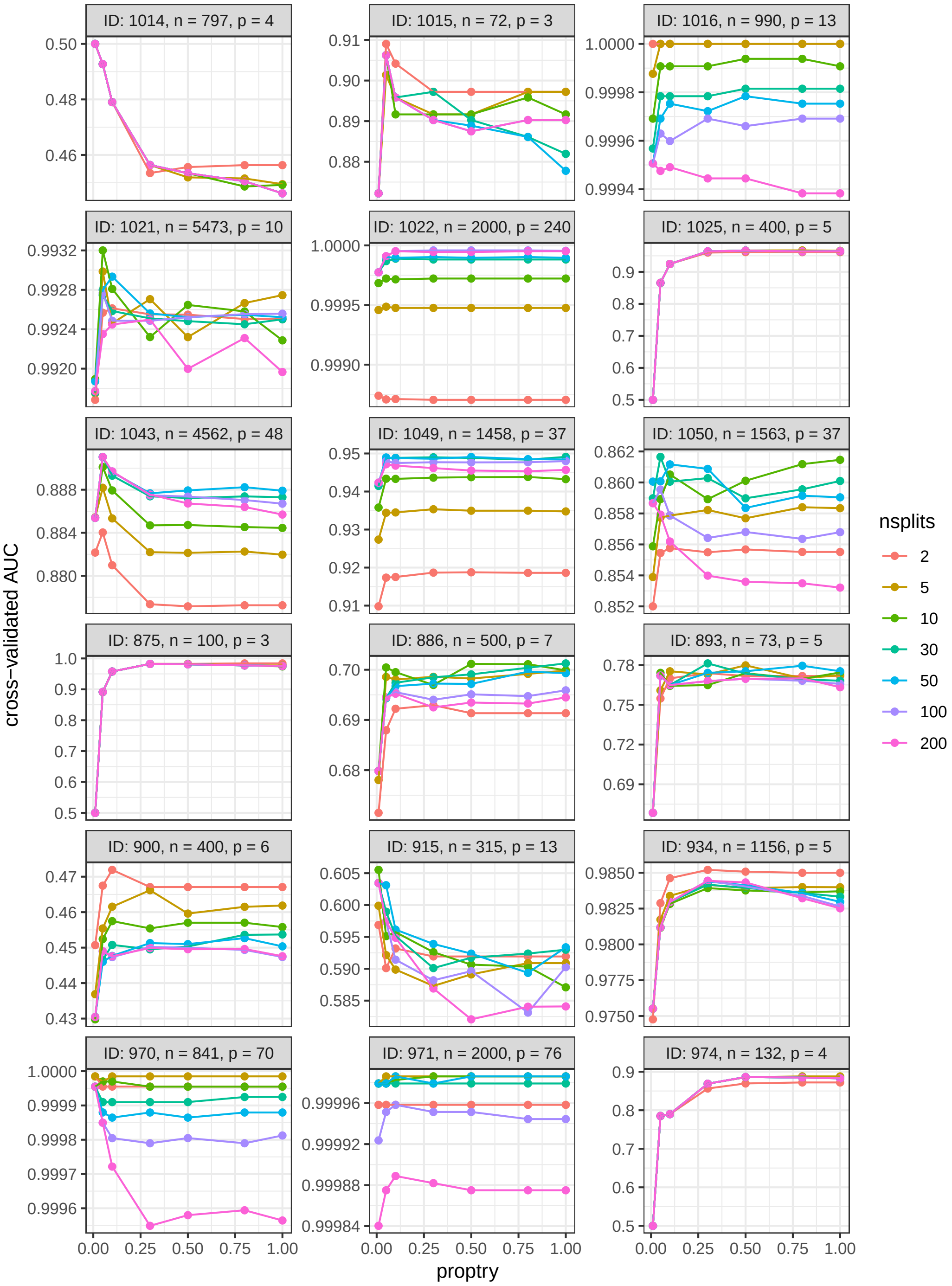

Fig. S2: Pre-study: Cross-validated AUC values obtained for the different parameter values. Each panel shows the results obtained for a particular data set. – II

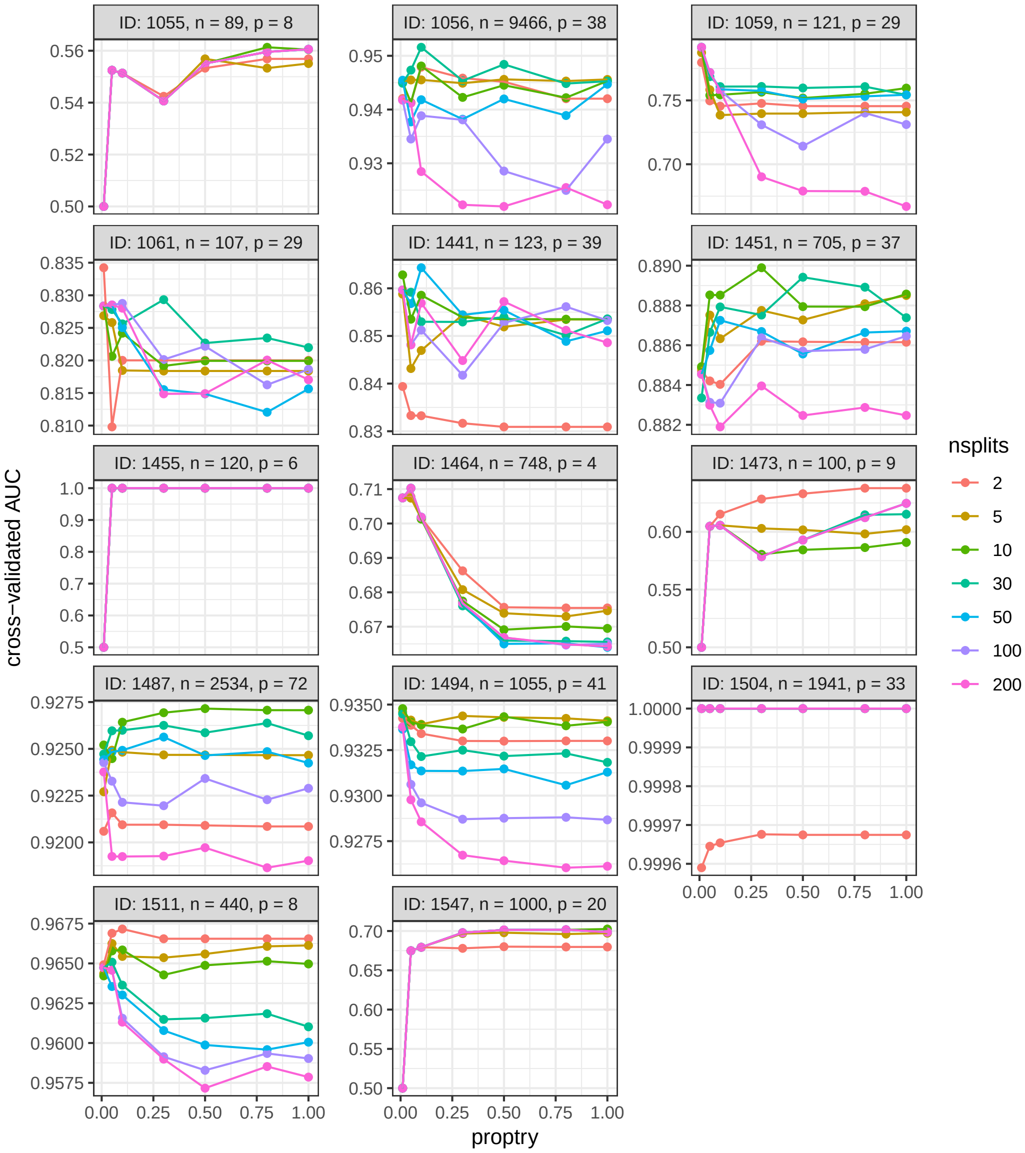

Fig. S3: Pre-study: Cross-validated AUC values obtained for different parameter values. Each panel shows the results obtained for a particular data set. – III

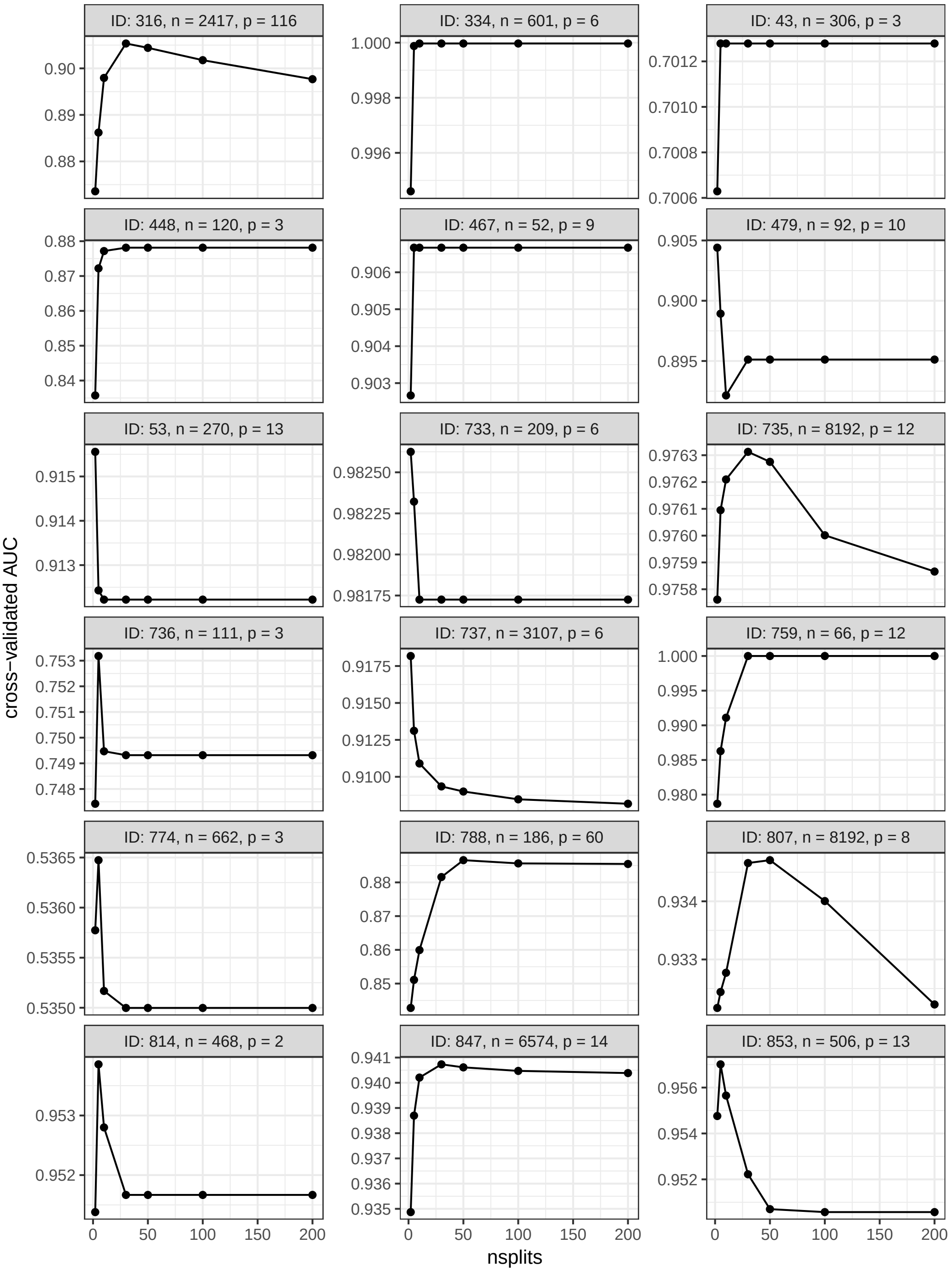

Fig. S4: Pre-study: Cross-validated AUC values obtained for different *nsplits* values. For each *nsplits* value considered, the plots show the maximum cross-validated AUC value obtained over the seven different values of *proptry*. Each panel shows the results obtained for a particular data set. – I

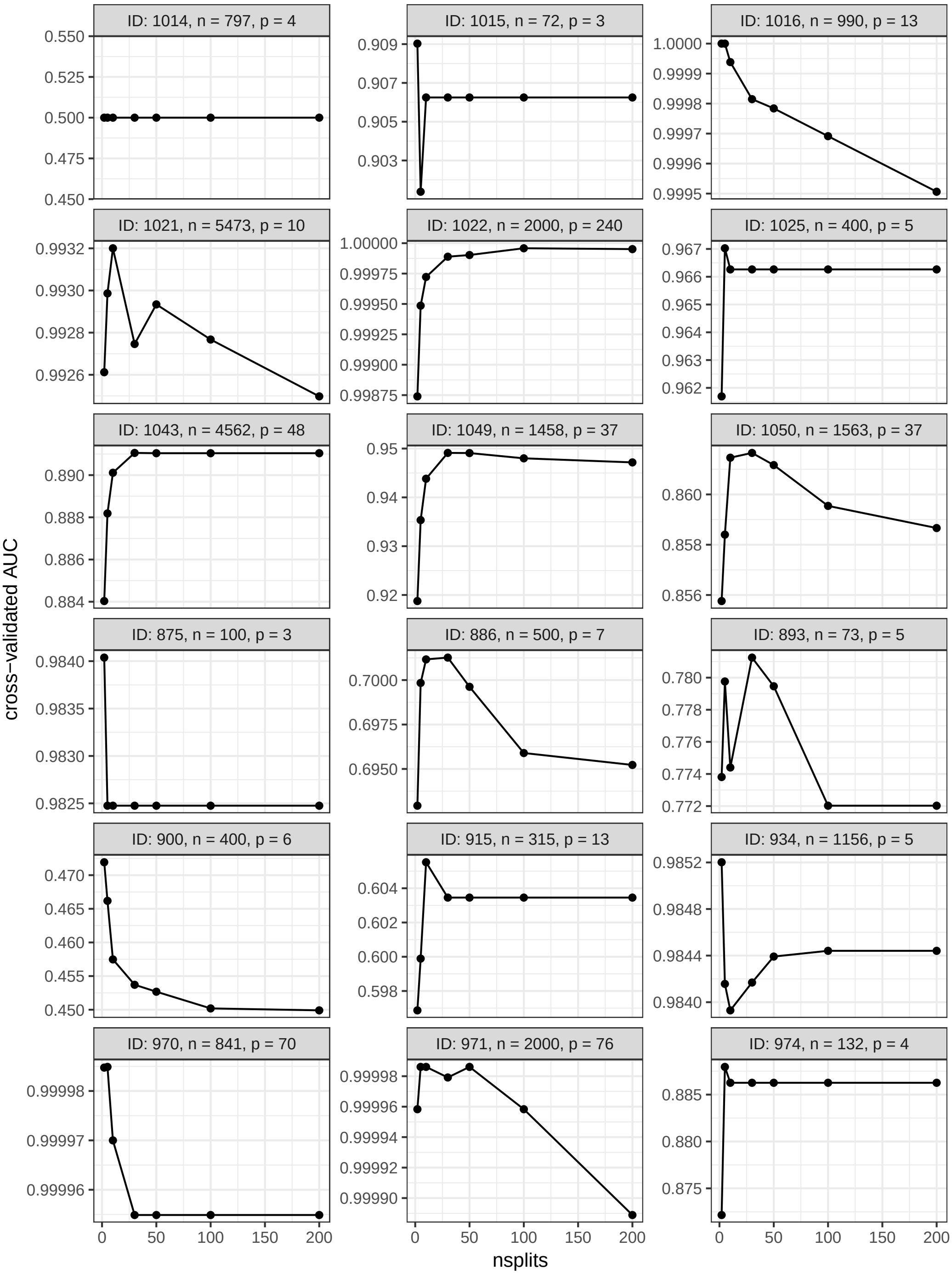

Fig. S5: Pre-study: Cross-validated AUC values obtained for different *nsplits* values. For each *nsplits* value considered, the plots show the maximum cross-validated AUC value obtained over the seven different values of *proptry*. Each panel shows the results obtained for a particular data set. – II

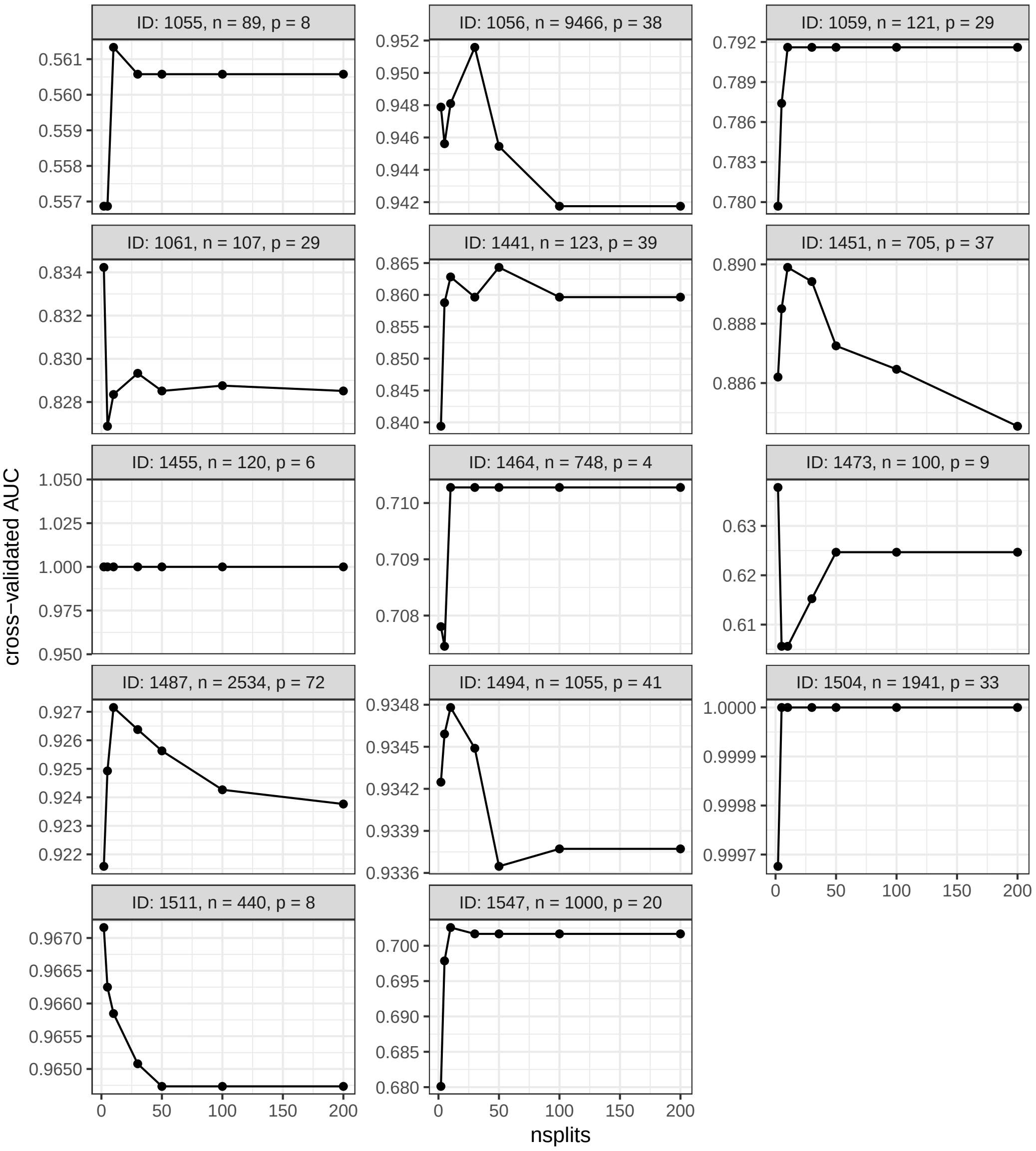

Fig. S6: Pre-study: Cross-validated AUC values obtained for different *nsplits* values. For each *nsplits* value considered, the plots show the maximum cross-validated AUC value obtained over the seven different values of *proptry*. Each panel shows the results obtained for a particular data set. – III

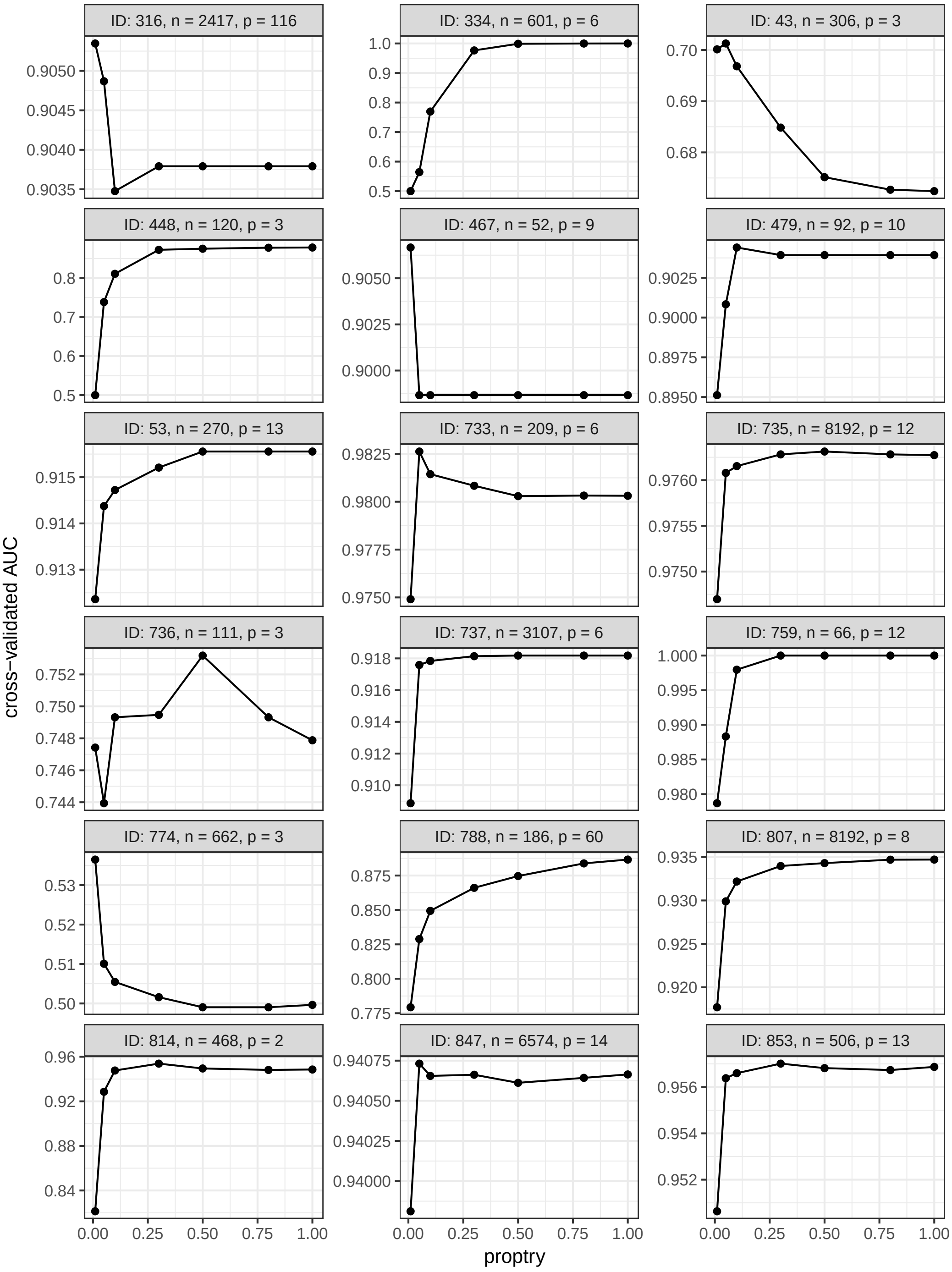

Fig. S7: Pre-study: Cross-validated AUC values obtained for different *proptry* values. For each *proptry* value considered, the plots show the maximum cross-validated AUC value obtained over the seven different values of *nsplits*. Each panel shows the results obtained for a particular data set. – I

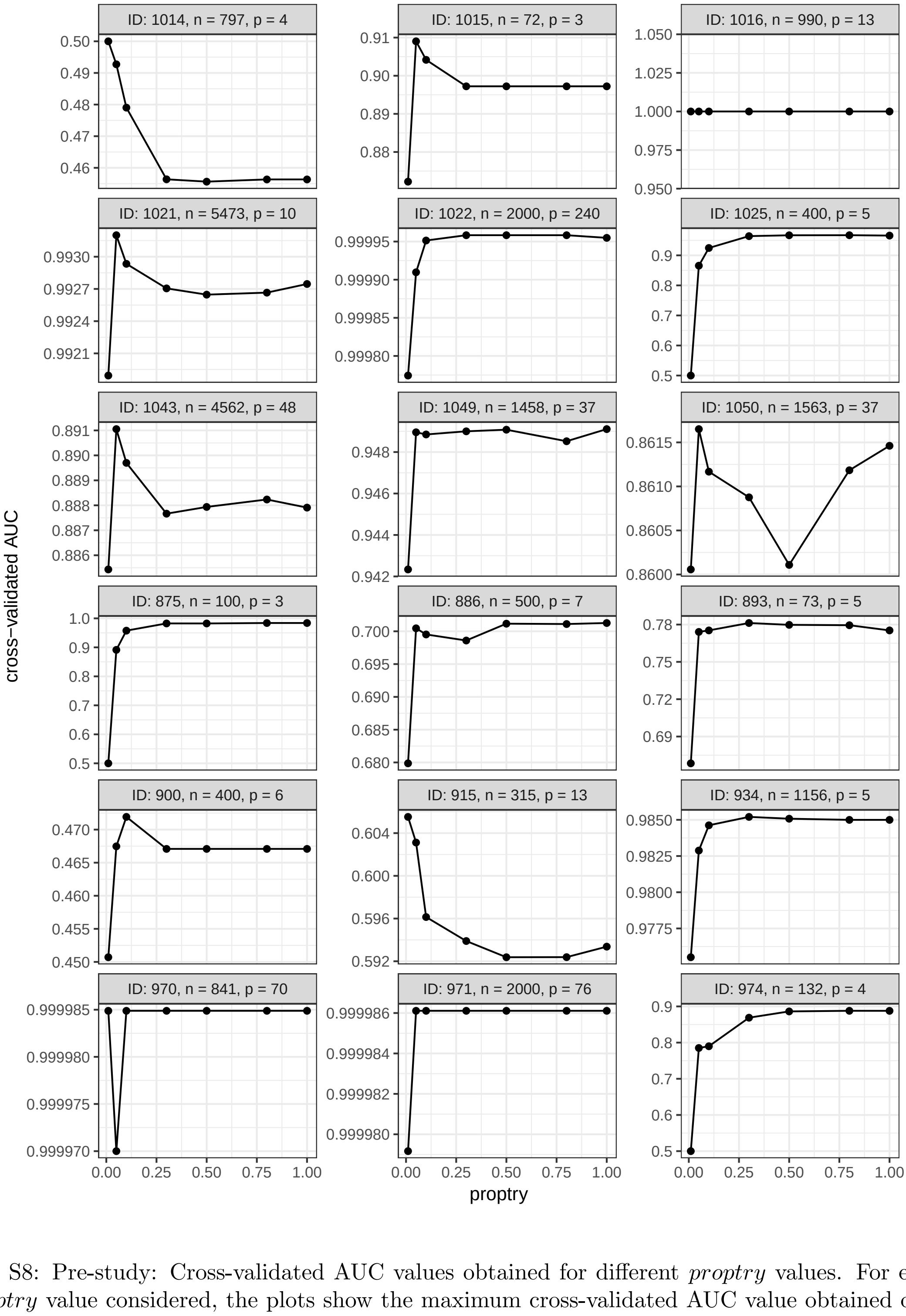

Fig. S8: Pre-study: Cross-validated AUC values obtained for different *proptry* values. For each *proptry* value considered, the plots show the maximum cross-validated AUC value obtained over the seven different values of *nsplits*. Each panel shows the results obtained for a particular data set. – II

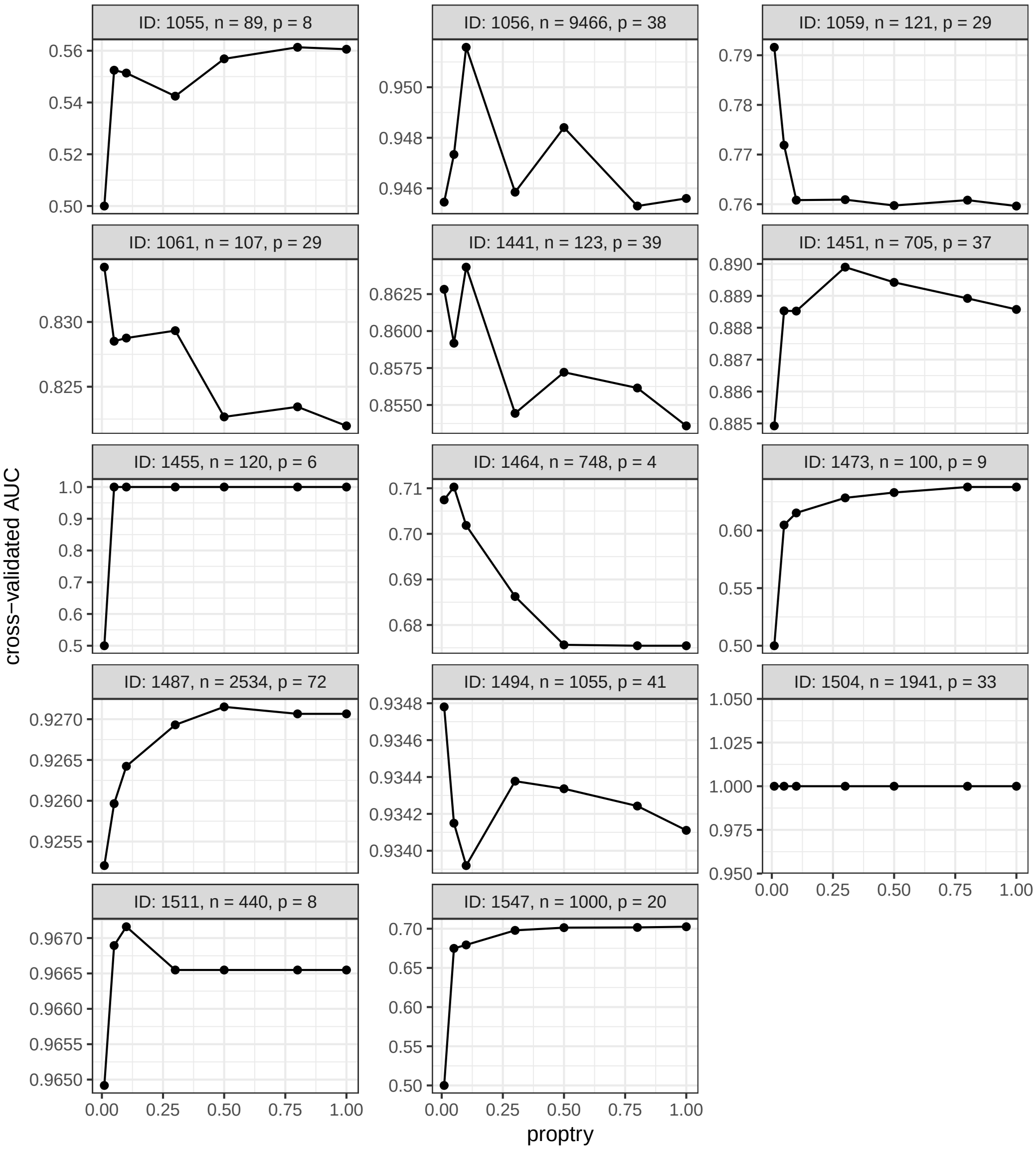

Fig. S9: Pre-study: Cross-validated AUC values obtained for different *proptry* values. For each *proptry* value considered, the plots show the maximum cross-validated AUC value obtained over the seven different values of *nsplits*. Each panel shows the results obtained for a particular data set. – III

C Data Set Specific Performances of RFextr1 and RFextr5 Compared to that of RF

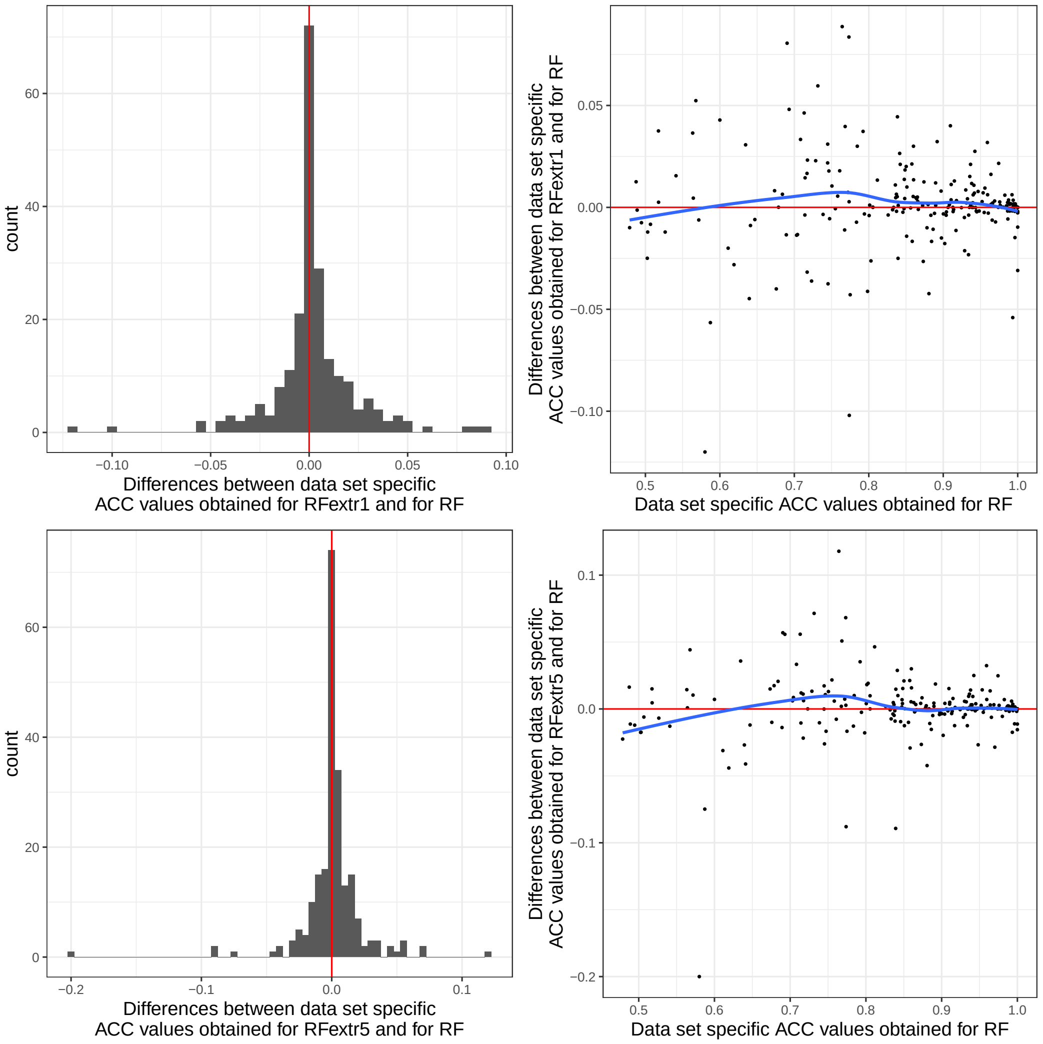

Fig. S10: Data set specific performances of RFextr1 and RFextr5 compared to that of RF. Left panels: Histograms of the differences between the data set specific ACC values obtained for RFextr1 / RFextr5 and for RF. The red lines indicate the zero line. Right panels: Scatter plot of the differences between the data set specific ACC values obtained for RFextr1 / RFextr5 and for RF against the data set specific ACC values obtained for RF. The blue lines represents loess fits. The red lines again indicate the zero line. The upper panels show the results obtained for RFextr1 and the lower panels those obtained for RFextr5.

D Influence of Sample Size and Number of Features on the Performance of RFextr1, RFextr5 and RF

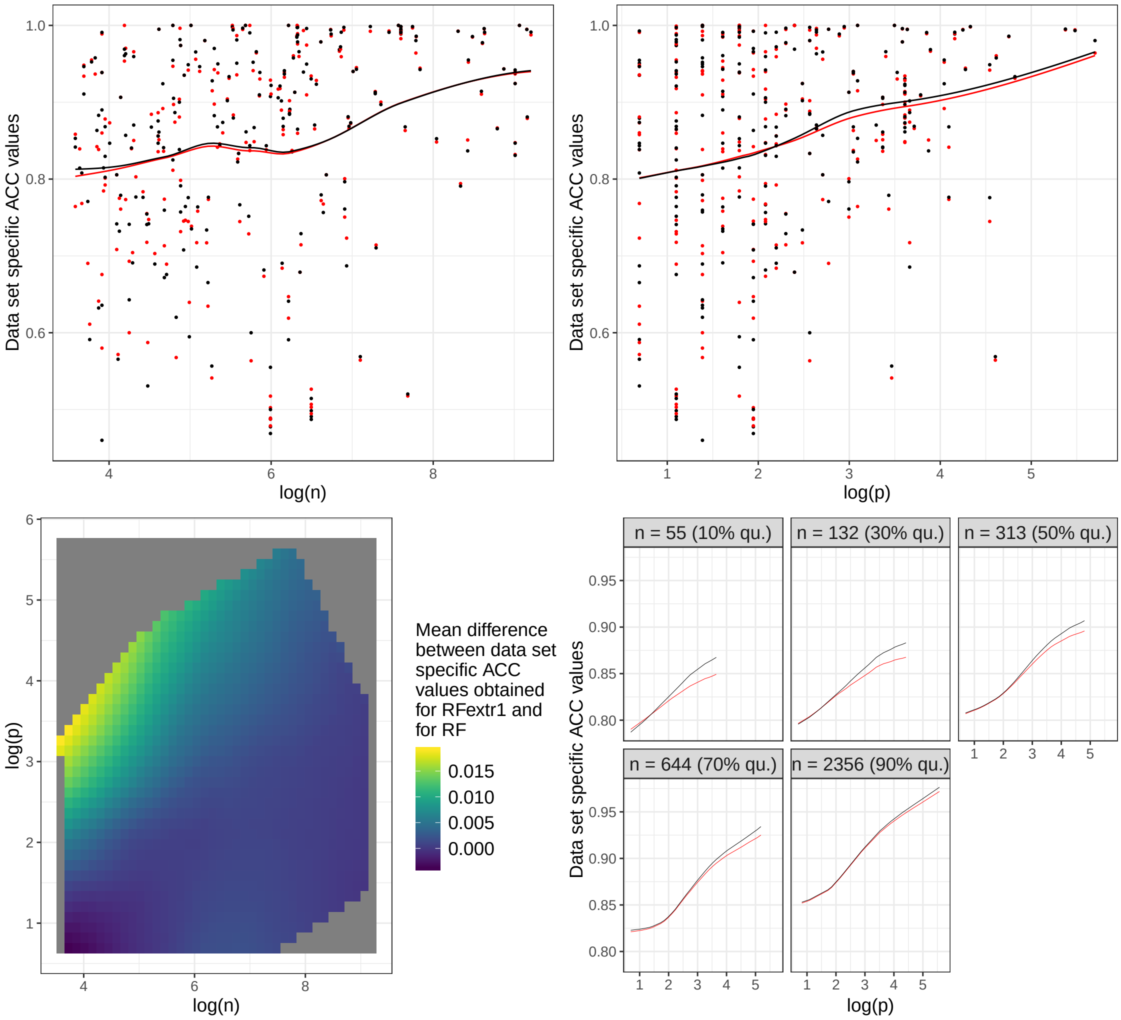

Fig. S11: Influence of sample size n and number of features p on the performance of RFextr1 and RF. Upper left / right panel: Data set specific ACC values obtained for RFextr1 and RF plotted against the logarithmized values of n and p. The lines show loess fits obtained for RFextr1 and RF, respectively. Lower left panel: Two-dimensional loess fit of the influences of the logarithmized values of n and p on the differences between the data set specific ACC values obtained for RFextr1 and for RF. Lower right panel: Cross sections of two-dimensional loess fits of the influences of the logarithmized values of n and p on the data set specific ACC values obtained for RFextr1 and RF, respectively. The cross sections were taken at different quantiles of the sample sizes of all data sets. Where applicable, in each plot the black lines show the results obtained for RFextr1 and the red lines those obtained for RF.

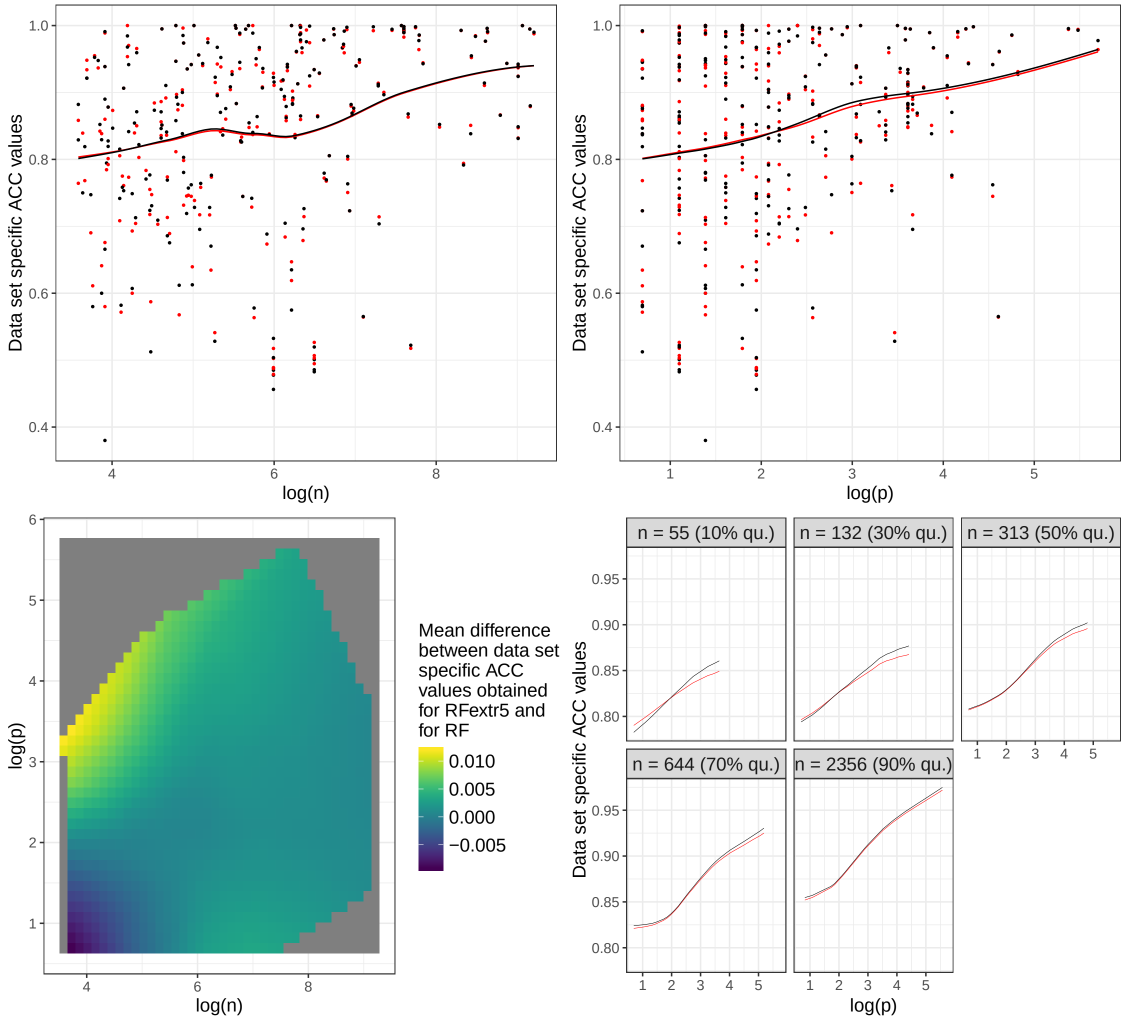

Fig. S12: Influence of sample size n and number of features p on the performance of RFextr5 and RF. Upper left / right panel: Data set specific ACC values obtained for RFextr5 and RF plotted against the logarithmized values of n and p. The lines show loess fits obtained for RFextr5 and RF, respectively. Lower left panel: Two-dimensional loess fit of the influences of the logarithmized values of n and p on the differences between the data set specific ACC values obtained for RFextr5 and for RF. Lower right panel: Cross sections of two-dimensional loess fits of the influences of the logarithmized values of n and p on the data set specific ACC values obtained for RFextr5 and RF, respectively. The cross sections were taken at different quantiles of the sample sizes of all data sets. Where applicable, in each plot the black lines show the results obtained for RFextr5 and the red lines those obtained for RF.

E Relationships Between the *mtry* Values Selected by RF, RFextr1, and RFextr5 and Various Quantities—Excluding Data Set '312'

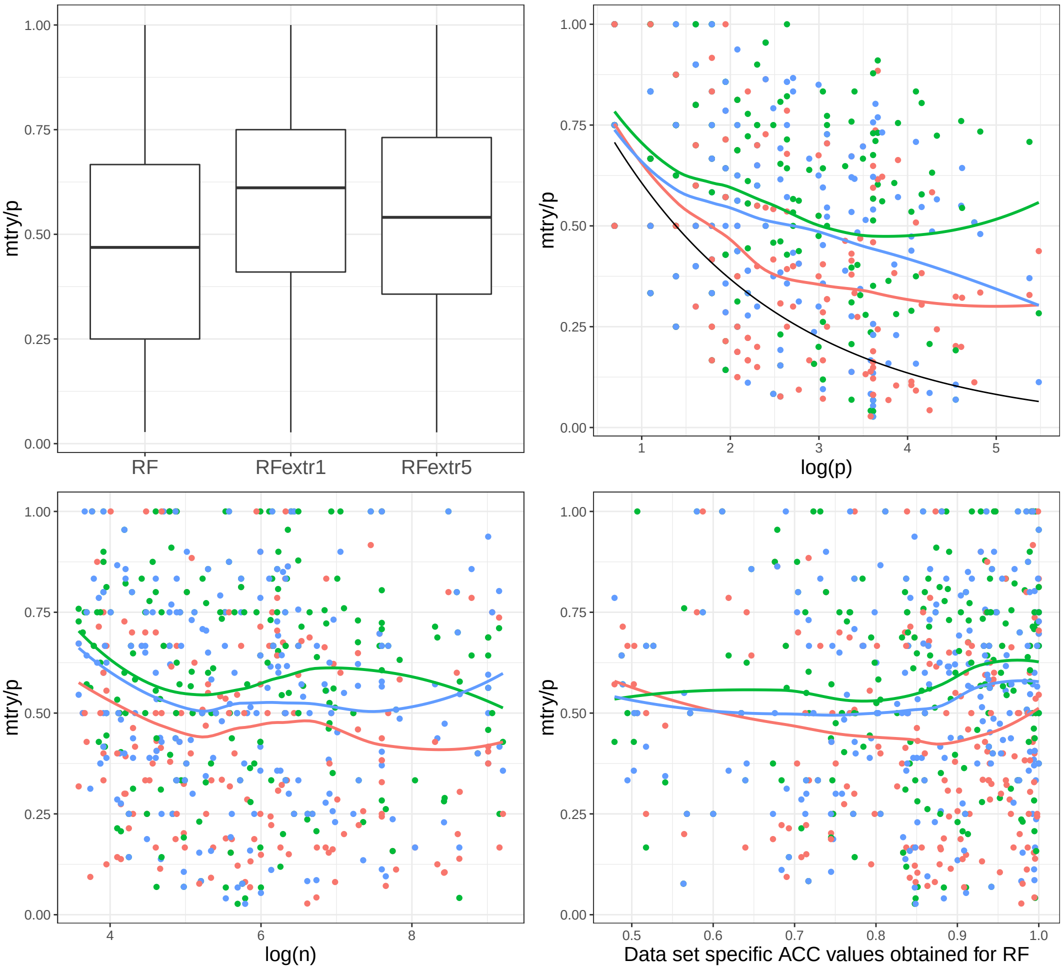

Fig. S13: Relationships between the *mtry* values selected by RF, RFextr1, and RFextr5 and various quantities—excluding data set '312'. Analogous to the analysis of the *proptry* values selected by DF, for each data set a single *mtry* value was considered in the plots. These data set specific *mtry* values were obtained by taking the median of the *mtry* values selected in the 10 training iterations of the two times repeated 5-fold stratified cross-validation. Upper left panel: Boxplots showing the *mtry* values divided by the numbers of features selected by RF, RFextr1, and RFextr5. Upper right / lower left / lower right panel: *mtry* values divided by the numbers of features plotted against the logarithmized values of the number of features (upper right panel), the logarithmized values of the sample size (lower left panel), and the data set specific ACC values obtained for RF (lower right panel). The black line in the upper right panel shows the $mtry/p$ values associated with the default choice $mtry = \sqrt{p}$. The different colors distinguish the different methods, where the results obtained for RF are shown in red, those obtained for RFextr1 in green and those obtained for RFextr5 in blue. The colored lines show loess fits.

F Relationships Between the *mtry* Values Selected by RF, RFextr1, and RFextr5 and the Logarithmized Values of the Sample Size Stratified According to the Numbers of Features in the Data Sets

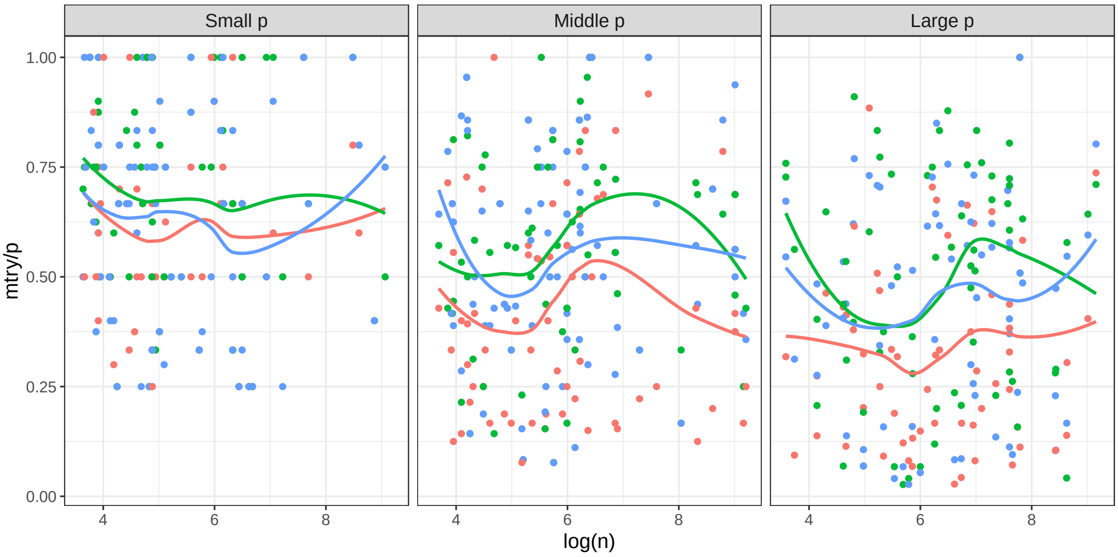

Fig. S14: Relationships between the *mtry* values selected by RF, RFextr1, and RFextr5 and the logarithmized values of the sample size stratified according to the numbers of features in the data sets. The data set specific *mtry* values shown in the plots were obtained in the same way as in the case of Figure S13. Left / middle / right panel: results obtained for the data sets with small ($p \leq 5$, 75 data sets), medium ($5 < p \leq 15$, 79 data sets), and large ($p > 15$, 66 data sets) numbers of features. The different colors distinguish the different methods, where the results obtained for RF are shown in red, those obtained for RFextr1 in green and those obtained for RFextr5 in blue. The lines shown loess fits.