Ovarian Cancer Histotypes: Report of Statistical Findings

Derek Chiu

2024-02-15

Contents

List of Figures

List of Tables

Preface

This report of statistical findings describes the classification of ovarian cancer histotypes using data from NanoString CodeSets.

Marina Pavanello conducted the initial exploratory data analysis, Cathy Tang implemented class imbalance techniques, Derek Chiu conducted the normalization and statistical analysis, and Lauren Tindale and Aline Talhouk are the project leads.

1. Introduction

Ovarian cancer has five major histotypes: high-grade serous carcinoma (HGSC), low-grade serous carcinoma (LGSC), endometrioid carcinoma (ENOC), mucinous carcinoma (MUC), and clear cell carcinoma (CCOC). A common problem with classifying these histotypes is that there is a class imbalance issue. HGSC dominates the distribution, commonly accounting for 70% of cases in many patient cohorts, while the other four histotypes are spread over the rest of the cases. Subsampling methods like up-sampling, down-sampling, and SMOTE can be used to mitigate this problem.

The supervised learning is performed under a consensus framework: we consider various classification algorithms and use evaluation metrics like accuracy, F1-score, Kappa, and G-mean to inform the decision of which methods to carry forward for prediction in confirmation and validation sets.

2. Methods

We use 5 classification algorithms and 4 subsampling methods across 500 repetitions in the supervised learning framework for the Training Set, CS1 and CS2. The pipeline was run using SLURM batch jobs submitted to a partition on a CentOS 7 server. Implementations of the techniques below were called from the splendid package.

- Classifiers:
  - Random Forest
  - SVM
  - Adaboost
  - Multinomial Regression Model with Ridge Penalty
  - Multinomial Regression Model with LASSO Penalty
- Subsampling:
  - None
  - Down-sampling
  - Up-sampling
  - SMOTE

3. Distributions

3.1 Full Data

The histotype distributions on the full data are shown below.

3.2 Training Sets

3.2.1 CS1 Training Set Generation

We use the reference method to normalize CS1 to CS3.

- CS1 reference set: duplicate samples from CS1
  - Samples = 16
  - Genes = 72
- CS3 reference set: corresponding samples in CS3 also found in CS1 reference set
  - Samples = 9
  - Genes = 72
- CS1 validation set: remaining CS1 samples with reference set removed
  - Samples = 270
  - Genes = 72

The final CS1 training set has 251 samples on 72 genes after normalization and keeping only the major histotypes of interest.

Table 3.1: All CodeSet Histotype Groups

Histotype Group	CS1	CS2	CS3
HGSC	120	643	1643
non-HGSC	166	220	583

Table 3.2: All CodeSet Major Reviewed Histotypes

Reviewed Histotype	CS1	CS2	CS3	CS1 %	CS2 %	CS3 %
CCOC	48	61	174	18.0	7.4	8.1
ENOC	60	32	232	22.5	3.9	10.8
HGSC	120	643	1643	44.9	78.5	76.1
LGSC	20	21	40	7.5	2.6	1.9
MUC	19	62	69	7.1	7.6	3.2

Table 3.3: All CodeSet Reviewed Histotypes

Reviewed Histotype	CS1	CS2	CS3
CARCINOMA-NOS	0	1	23
CCOC	48	61	174
CTRL	0	12	0
ENOC	60	32	232
HGSC	120	643	1643
LGSC	20	21	40
MBOT	0	19	3
MIXED (ENOC/CCOC)	0	0	1
MIXED (ENOC/LGSC)	0	0	1
MIXED (HGSC/CCOC)	0	0	1
MMMT	0	0	28
MUC	19	62	69
Other/Exclude	0	0	8
SBOT	19	12	2
serous LMP	0	0	1

Table 3.4: CS1 Histotypes

CodeSet	Reviewed Histotype	n
CS1	CCOC	48
CS1	ENOC	60
CS1	HGSC	120
CS1	LGSC	20
CS1	MUC	19
CS1	SBOT	19

Table 3.5: CS2 Histotypes

CodeSet	Reviewed Histotype	n
CS2	CARCINOMA-NOS	1
CS2	CCOC	61
CS2	CTRL	12
CS2	ENOC	32
CS2	HGSC	643
CS2	LGSC	21
CS2	MBOT	19
CS2	MUC	62
CS2	SBOT	12

Table 3.6: CS3 Histotypes

CodeSet	Reviewed Histotype	n
CS3	CARCINOMA-NOS	23
CS3	CCOC	174
CS3	ENOC	232
CS3	HGSC	1643
CS3	LGSC	40
CS3	MBOT	3
CS3	MIXED (ENOC/CCOC)	1
CS3	MIXED (ENOC/LGSC)	1
CS3	MIXED (HGSC/CCOC)	1
CS3	MMMT	28
CS3	MUC	69
CS3	Other/Exclude	8
CS3	SBOT	2
CS3	serous LMP	1

Table 3.7: Common Summary ID CodeSet Histotypes

Reviewed Histotype	CS1	CS2	CS3
CCOC	3	4	9
ENOC	4	4	9
HGSC	55	62	94
LGSC	7	5	8
MUC	7	5	11

Table 3.8: CS1 Training Set Histotypes

Histotype	n	%
CCC	57	18.8%
ENOCa	59	19.4%
HGSC	156	51.3%
LGSC	16	5.3%
MUC	16	5.3%

Table 3.9: CS2 Training Set Histotypes

Histotype	n	%
CCOC	68	7.2%
ENOC	30	3.2%
HGSC	757	80.1%
LGSC	29	3.1%
MUC	61	6.5%

Table 3.10: All Common Samples Histotype Distribution

revHist	CS1	CS2	CS3
CCOC	3	4	3
ENOC	4	4	3
HGSC	53	56	68
LGSC	7	5	4
MUC	7	5	5

3.2.2 CS2 Training Set Generation

We use the pool method to normalize CS2 to CS3 so we can be consistent with the PrOType normalization when there are available pools.

- CS2 pools:
  - Samples = 12 (Pool 1 = 4, Pool 2 = 4, Pool 3 = 4)
  - Genes = 365
- CS3 pools:
  - Samples = 22 (Pool 1 = 12, Pool 2 = 5, Pool 3 = 5)
  - Genes = 513
- CS2 validation set: CS2 samples with pools removed
  - Samples = 879
  - Genes = 365

The final CS2 training set has 819 samples on 136 (common) genes after normalization and keeping only the major histotypes of interest.

Table 3.11: Distinct Common Samples Histotype Distribution

revHist	CS1	CS2	CS3
CCOC	3	3	3
ENOC	3	3	3
HGSC	51	51	51
LGSC	4	4	4
MUC	5	5	5

Table 3.12: Distinct Common CS2 and CS3 Samples Histotype Distribution

revHist	CS2	CS3
CCOC	3	3
ENOC	3	3
HGSC	71	71
LGSC	4	4
MUC	5	5

Table 3.13: Common Samples Across Sites Histotype Distribution

revHist	AOC	USC	Vancouver
CCOC	3	3	3
ENOC	3	3	3
HGSC	13	13	26
LGSC	2	2	2
MUC	3	3	3

Table 3.14: Distinct Common Samples Across Sites Histotype Distribution

revHist	AOC	USC	Vancouver
CCOC	3	3	3
ENOC	3	3	3
HGSC	13	13	13
LGSC	2	2	2
MUC	3	3	3

Table 3.15: CS3/CS4/CS5 Common Samples Histotype Distribution

revHist	CS3	CS4	CS5
HGSC	47	47	47
NA	26	26	26

Table 3.16: CS3/CS4/CS5 Pools Distribution

Pool	CS3	CS4	CS5
Pool1	12	4	4
Pool2	5	4	4
Pool3	5	4	4
Pool4	NA	1	1
Pool5	NA	1	1
Pool6	NA	1	0
Pool7	NA	1	1
Pool8	NA	1	1
Pool9	NA	1	1
Pool10	NA	1	1
Pool11	NA	1	1

Table 3.17: Pre-QC Training Set Histotype Distribution by CodeSet

Variable	Levels	CS1	CS2	CS3	Total
Histotype	HGSC	120 (45%)	643 (79%)	515 (92%)	1278 (78%)
	CCOC	48 (18%)	61 (7%)	11 (2%)	120 (7%)
	ENOC	60 (22%)	32 (4%)	11 (2%)	103 (6%)
	MUC	19 (7%)	62 (8%)	12 (2%)	93 (6%)
	LGSC	20 (7%)	21 (3%)	9 (2%)	50 (3%)
Total	N (%)	267 (16%)	819 (50%)	558 (34%)	1644 (100%)

Table 3.18: Full Training Set Histotype Distribution by CodeSet

Variable	Levels	CS1	CS2	CS3	Total
Histotype	HGSC	116 (48%)	623 (80%)	475 (94%)	1214 (79%)
	CCOC	44 (18%)	54 (7%)	8 (2%)	106 (7%)
	ENOC	55 (23%)	27 (3%)	8 (2%)	90 (6%)
	MUC	15 (6%)	59 (8%)	9 (2%)	83 (5%)
	LGSC	14 (6%)	19 (2%)	6 (1%)	39 (3%)
Total	N (%)	244 (16%)	782 (51%)	506 (33%)	1532 (100%)

Table 3.19: Histotype Distribution by CodeSet/Datasets

Variable	Levels	CS1 All	CS2 All	Confirmation	Validation
Histotype	HGSC	119 (46%)	642 (79%)	422 (66%)	674 (74%)
	CCOC	47 (18%)	60 (7%)	75 (12%)	80 (9%)
	ENOC	58 (22%)	30 (4%)	106 (16%)	108 (12%)
	MUC	18 (7%)	61 (8%)	27 (4%)	26 (3%)
	LGSC	18 (7%)	20 (2%)	13 (2%)	18 (2%)
Total	N (%)	260 (10%)	813 (31%)	643 (25%)	906 (35%)

Table 3.20: Number of failed sampled by CodeSet

CS1	CS2	CS3
8	32	8

3.3 Common Samples

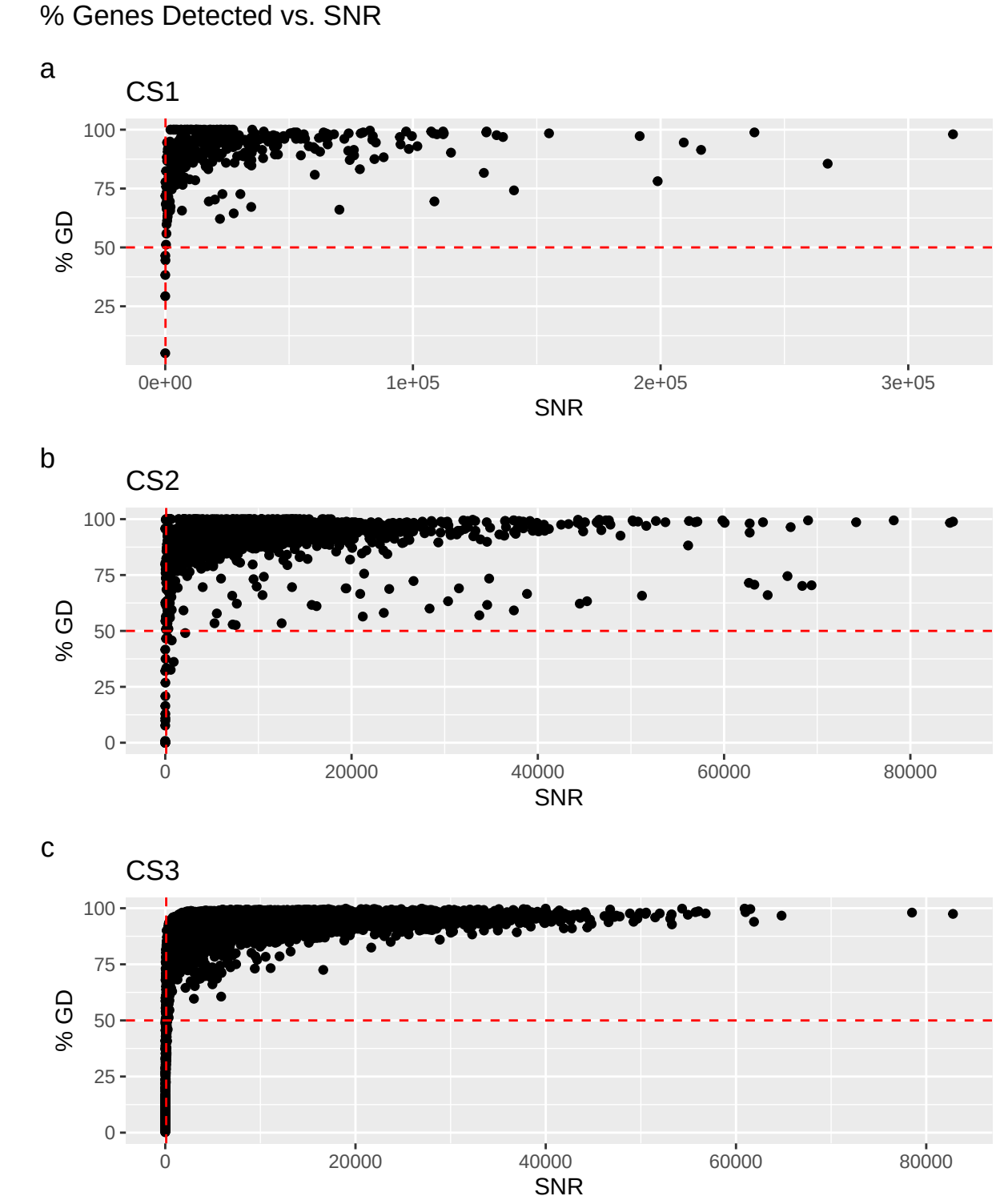
3.4 Histotypes in Classifier Data

3.5 Quality Control

3.5.1 Failed Samples

3.5.2 %GD vs. SNR

\begin{figure}[H]

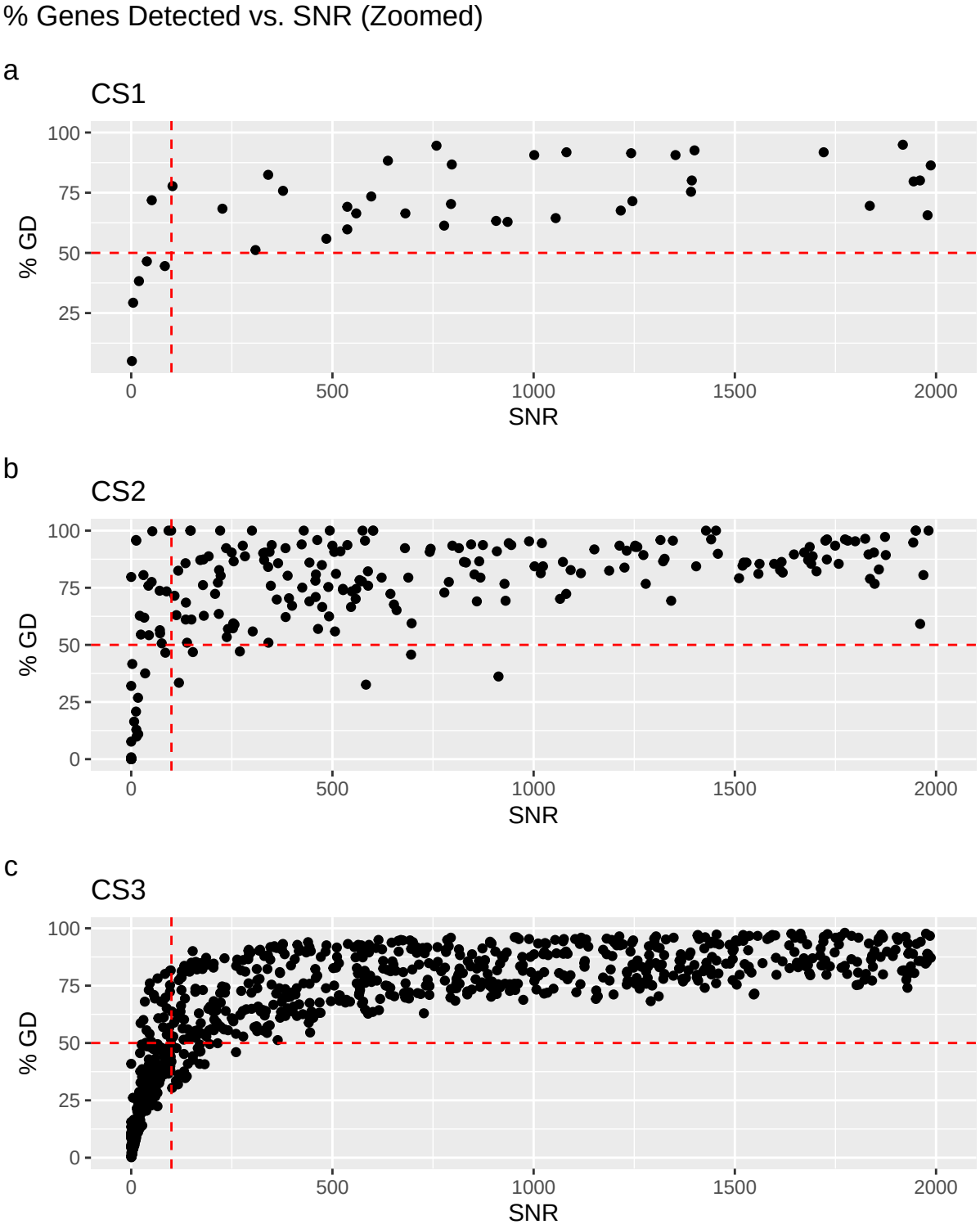


{

}

\caption{% Genes Detected vs. Signal to Noise Ratio} \end{figure}

\begin{figure}[H]



{

}

\caption{% Genes Detected vs. Signal to Noise Ratio (Zoomed)} \end{figure}

4. Results

We show internal validation summaries for the combined classifier training set, as well as the CS1 and CS2 sets with duplicates included. The F1-scores, kappa, and G-mean are the measures of interest. Algorithms are sorted by descending value based on the overallaccuracy of the training set. The point ranges show the median, 5th and 95th percentiles, coloured by subsampling methods.

4.1 Training Set

4.1.1 Accuracy

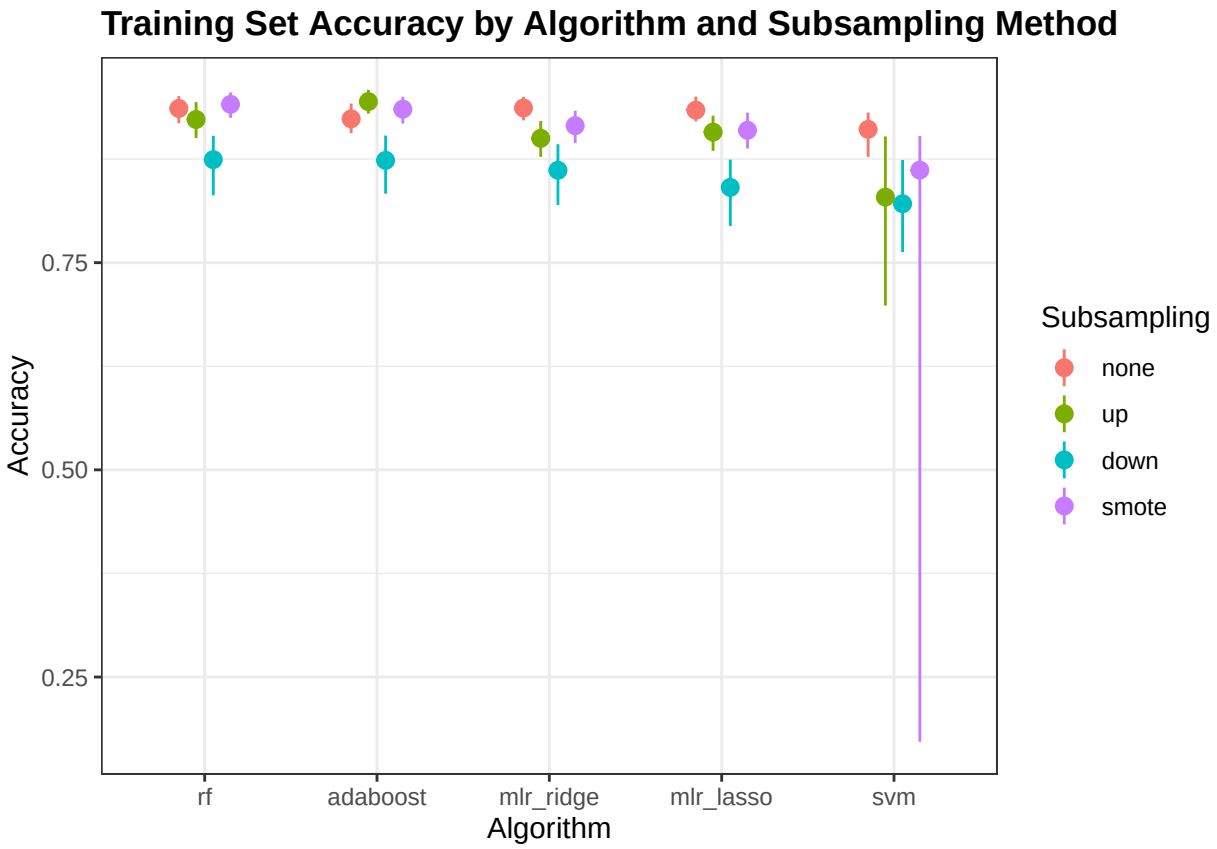


Figure 4.1: Training Set Accuracy

Table 4.1: Training Set Accuracy by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.936	0.924	0.937	0.934	0.911
up	0.923	**0.944**	0.9	0.908	0.829
down	0.874	0.873	0.862	0.841	0.821
smote	0.941	0.935	0.915	0.91	0.862

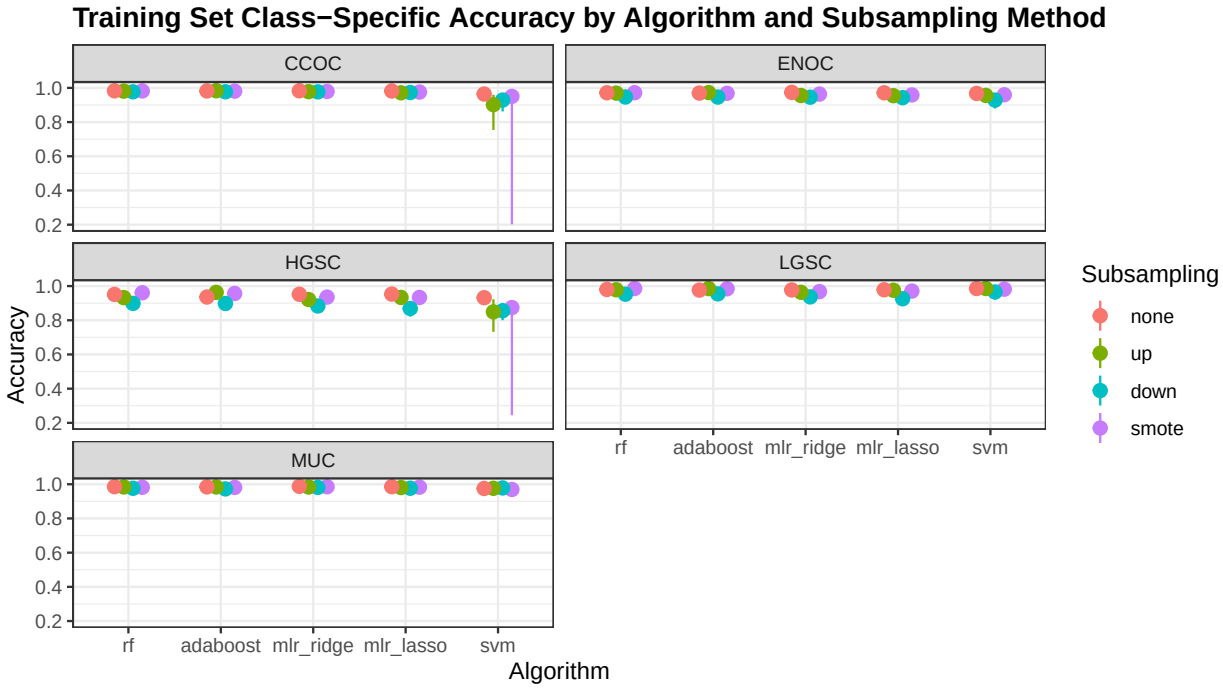


Figure 4.2: Training Set Class-Specific Accuracy

Table 4.2: Training Set Class-Specific Accuracy by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.983	0.982	0.983	0.982	0.964
none	ENOC	0.971	0.97	0.974	0.971	0.968
none	HGSC	0.951	0.936	0.952	0.953	0.932
none	LGSC	0.98	0.976	0.977	0.978	0.986
none	MUC	0.986	0.984	**0.988**	0.986	0.975
up	CCOC	0.981	0.983	0.978	0.971	0.902
up	ENOC	0.969	0.972	0.955	0.954	0.956
up	HGSC	0.932	0.963	0.921	0.934	0.849
up	LGSC	0.978	0.985	0.964	0.976	0.986
up	MUC	0.984	0.984	0.984	0.981	0.976
down	CCOC	0.977	0.977	0.977	0.973	0.929
down	ENOC	0.947	0.946	0.946	0.943	0.929
down	HGSC	0.899	0.898	0.884	0.869	0.856
down	LGSC	0.953	0.955	0.937	0.926	0.965
down	MUC	0.976	0.973	0.982	0.977	0.978
smote	CCOC	0.982	0.981	0.979	0.976	0.95
smote	ENOC	0.972	0.968	0.964	0.959	0.96
smote	HGSC	0.961	0.957	0.936	0.934	0.874
smote	LGSC	0.986	0.984	0.968	0.971	0.982
smote	MUC	0.982	0.982	0.985	0.983	0.97

Table 4.3: Training Set Macro-Averaged F1-Score by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.769	0.726	0.766	0.779	0.755
up	0.723	**0.827**	0.772	0.761	0.688
down	0.732	0.73	0.727	0.699	0.672
smote	0.826	0.82	0.789	0.777	0.614

4.1.2 F1-Score

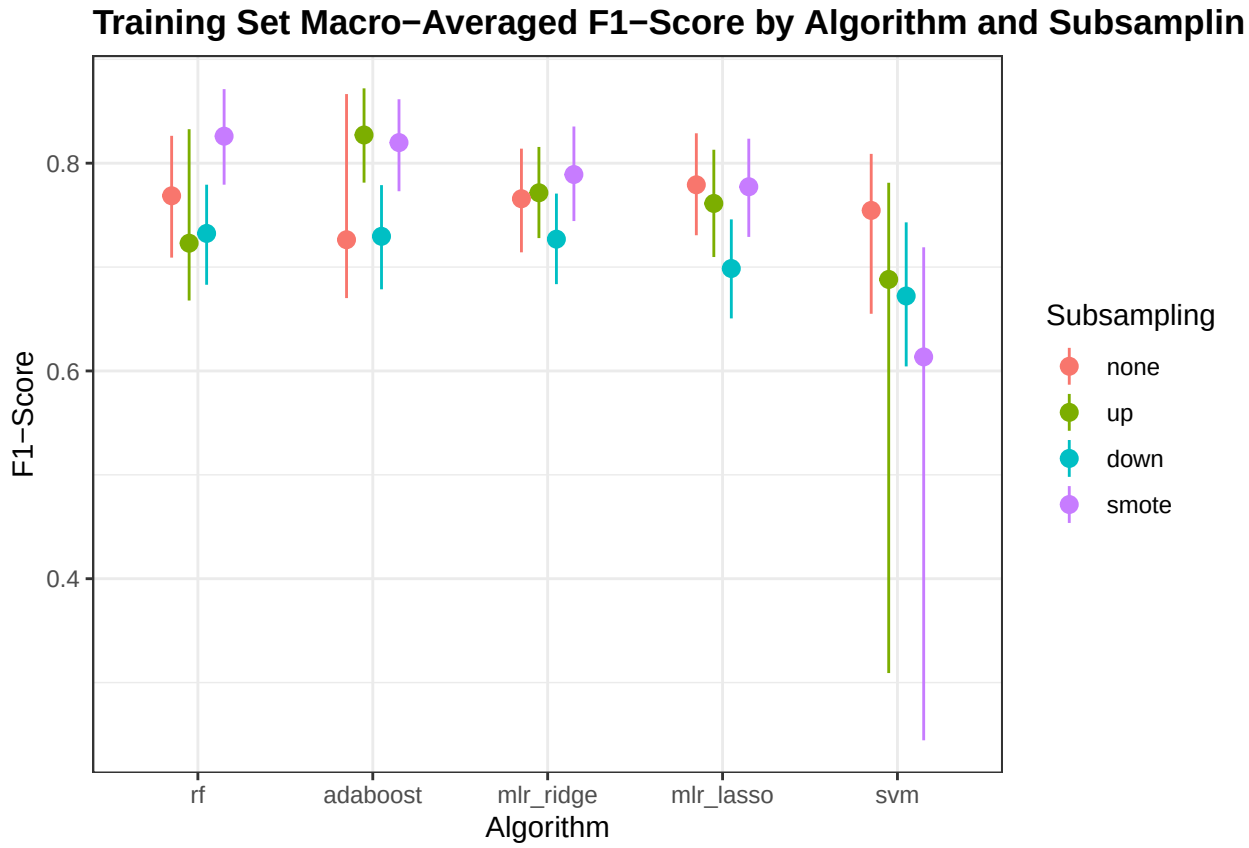


Figure 4.3: Training Set F1-Score

Table 4.4: Training Set Class-Specific F1-Score by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.871	0.857	0.87	0.861	0.735
none	ENOC	0.73	0.688	0.759	0.747	0.712
none	HGSC	0.97	0.961	0.97	0.971	0.958
none	LGSC	0.417	0.182	0.364	0.476	0.667
none	MUC	0.863	0.851	0.885	0.865	0.711
up	CCOC	0.849	0.872	0.844	0.794	0.491
up	ENOC	0.677	0.769	0.676	0.648	0.646
up	HGSC	0.959	**0.977**	0.948	0.957	0.905
up	LGSC	0.267	0.667	0.545	0.6	0.645
up	MUC	0.846	0.857	0.857	0.821	0.731
down	CCOC	0.833	0.838	0.833	0.811	0.632
down	ENOC	0.63	0.628	0.644	0.615	0.566
down	HGSC	0.932	0.932	0.921	0.91	0.901
down	LGSC	0.47	0.477	0.421	0.375	0.514
down	MUC	0.8	0.776	0.838	0.794	0.776
smote	CCOC	0.865	0.861	0.853	0.825	0.58
smote	ENOC	0.761	0.738	0.718	0.685	0.571
smote	HGSC	0.975	0.972	0.958	0.957	0.924
smote	LGSC	0.706	0.71	0.567	0.579	0.429
smote	MUC	0.836	0.831	0.862	0.843	0.622

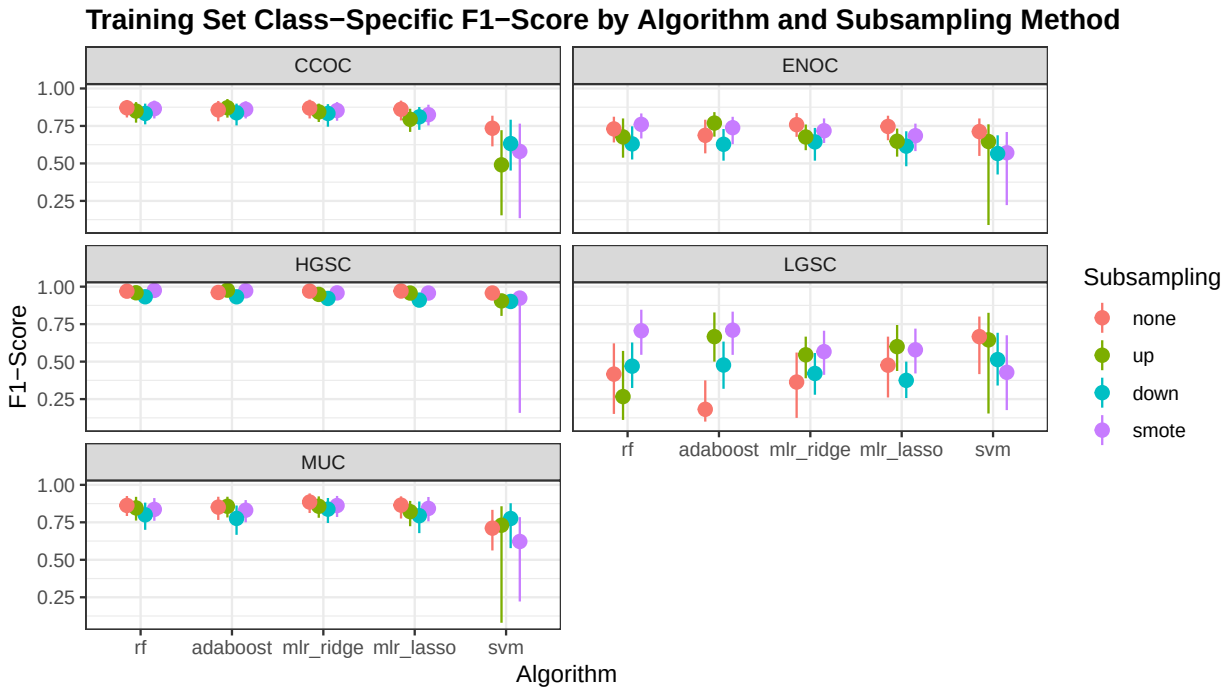


Figure 4.4: Training Set Class-Specific F1-Score

Table 4.5: Training Set Kappa by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.806	0.758	0.812	0.809	0.737
up	0.751	**0.841**	0.754	0.754	0.582
down	0.702	0.701	0.682	0.643	0.605
smote	0.833	0.823	0.782	0.768	0.561

4.1.3 Kappa

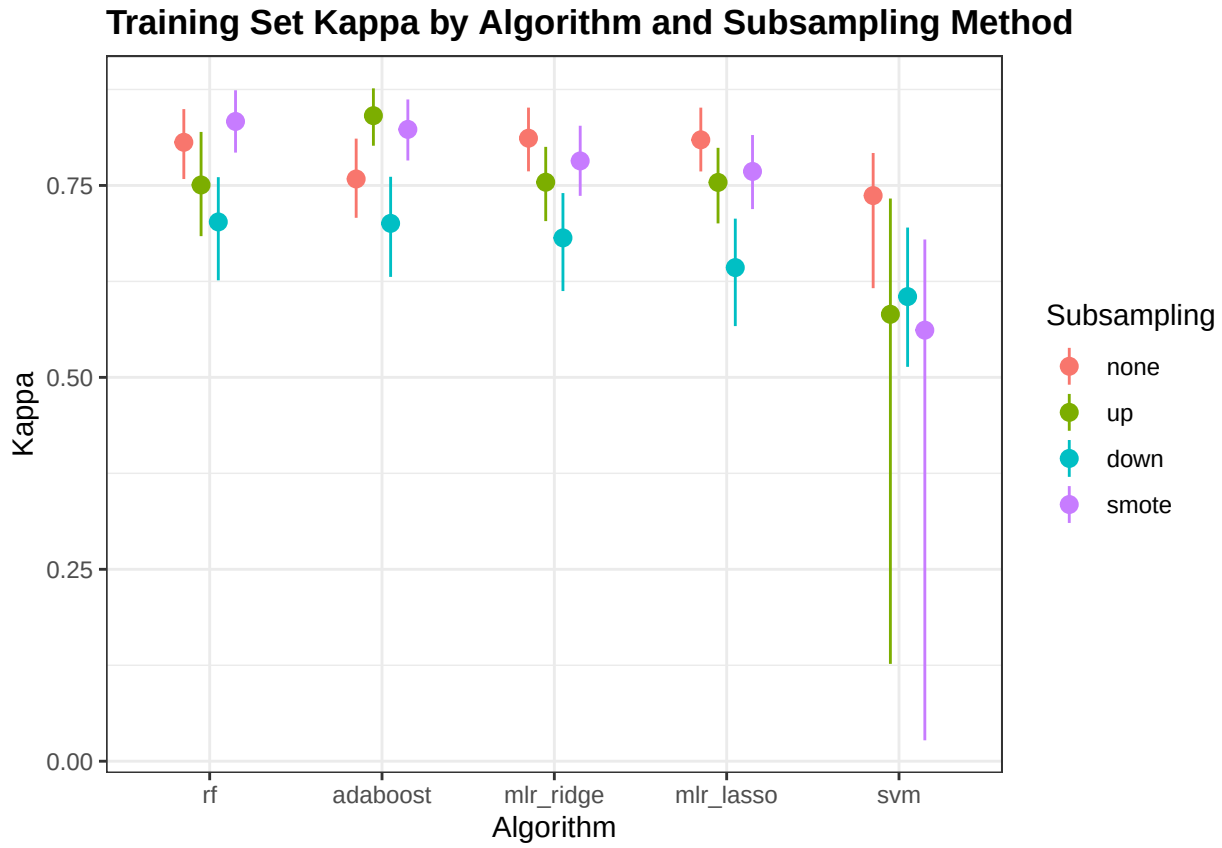


Figure 4.5: Training Set Kappa

Table 4.6: Training Set Class-Specific Kappa by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.862	0.848	0.859	0.852	0.714
none	ENOC	0.714	0.671	0.746	0.733	0.695
none	HGSC	0.84	0.783	0.845	0.852	0.781
none	LGSC	0.406	0.125	0.353	0.466	0.659
none	MUC	0.856	0.844	0.879	0.856	0.7
up	CCOC	0.839	0.862	0.832	0.777	0.438
up	ENOC	0.659	0.754	0.654	0.624	0.624
up	HGSC	0.765	**0.887**	0.784	0.807	0.599
up	LGSC	0.23	0.66	0.529	0.59	0.634
up	MUC	0.839	0.849	0.848	0.811	0.718
down	CCOC	0.822	0.825	0.82	0.797	0.599
down	ENOC	0.605	0.601	0.616	0.585	0.53
down	HGSC	0.729	0.73	0.701	0.667	0.64
down	LGSC	0.45	0.459	0.396	0.349	0.496
down	MUC	0.786	0.762	0.828	0.782	0.764
smote	CCOC	0.856	0.851	0.842	0.812	0.556
smote	ENOC	0.745	0.722	0.699	0.663	0.548
smote	HGSC	0.881	0.87	0.817	0.811	0.569
smote	LGSC	0.699	0.701	0.551	0.565	0.423
smote	MUC	0.828	0.822	0.855	0.834	0.608

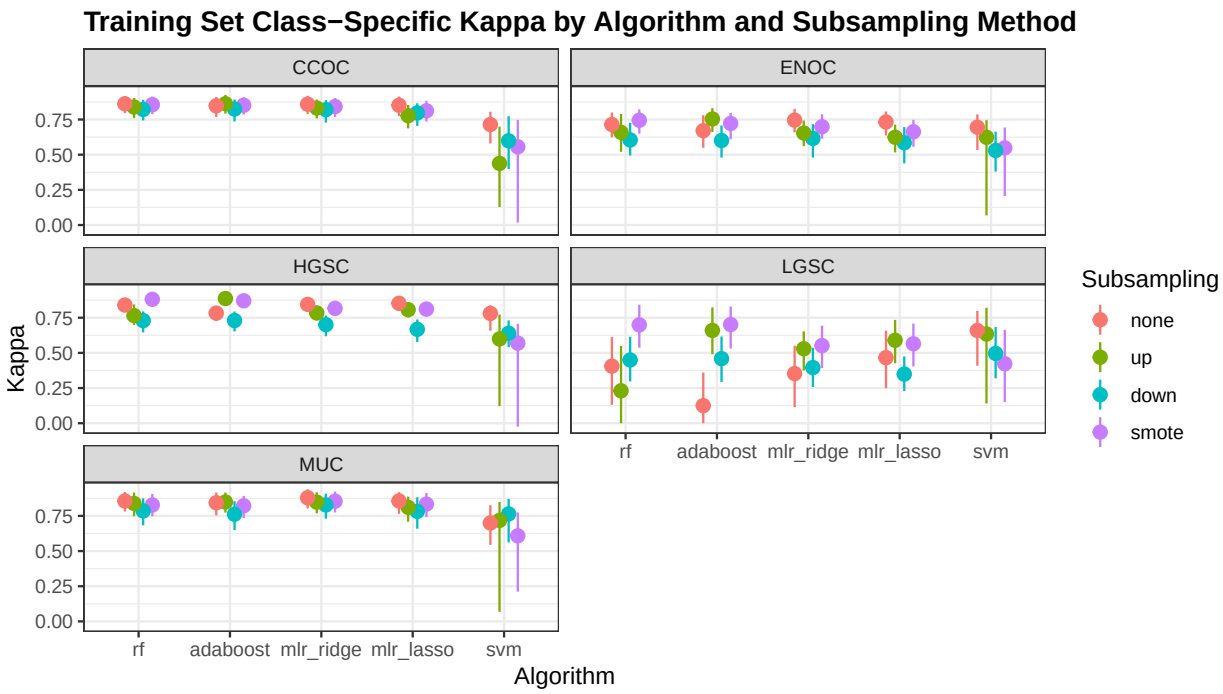


Figure 4.6: Training Set Class-Specific Kappa

Table 4.7: Training Set G-mean by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.657	0.481	0.663	0.72	0.695
up	0.528	0.799	**0.871**	0.8	0.706
down	0.849	0.846	0.861	0.842	0.788
smote	0.822	0.837	0.862	0.838	0.501

4.1.4 G-mean

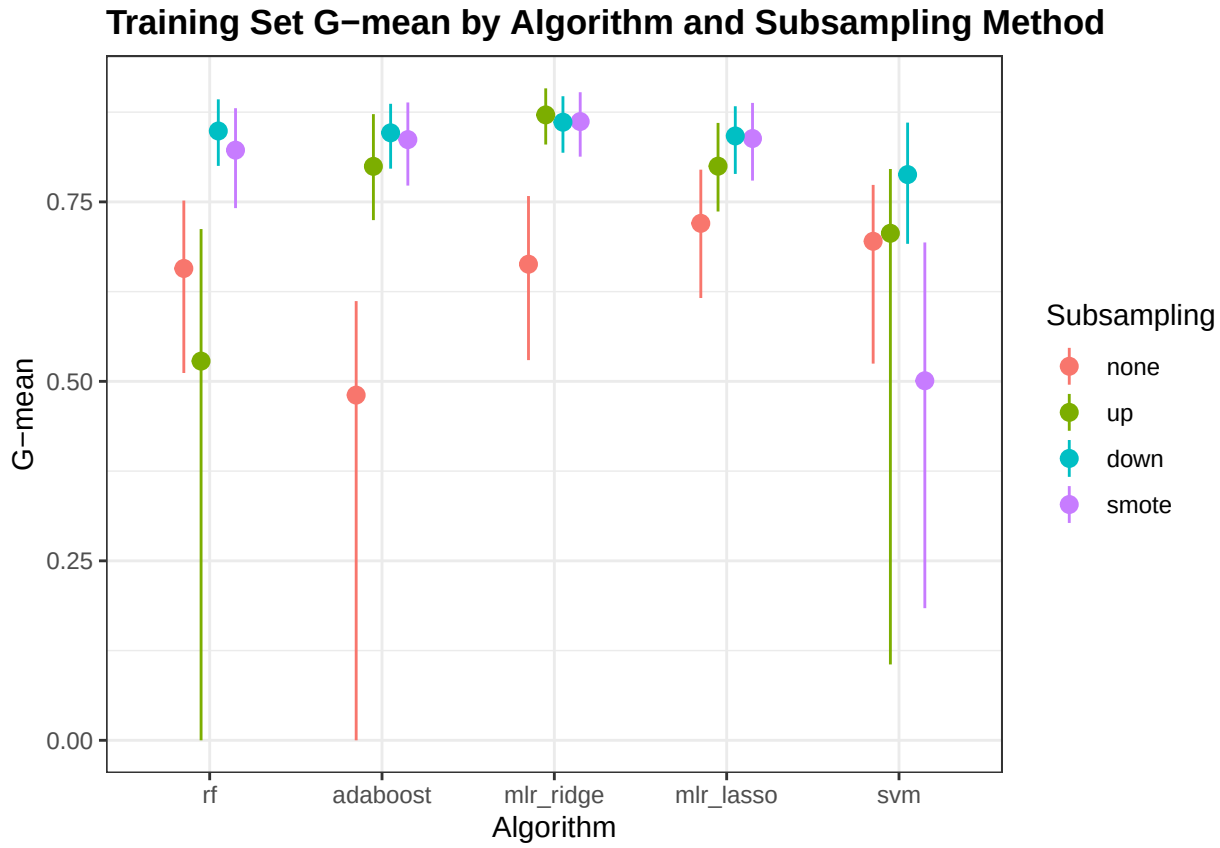


Figure 4.7: Training Set G-mean

Table 4.8: Training Set Class-Specific G-mean by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.907	0.888	0.904	0.912	0.856
none	ENOC	0.812	0.758	0.844	0.84	0.828
none	HGSC	0.886	0.838	0.895	0.909	0.866
none	LGSC	0.527	0.267	0.5	0.63	0.755
none	MUC	0.915	0.9	0.93	0.919	0.753
up	CCOC	0.876	0.91	0.925	0.892	0.873
up	ENOC	0.737	0.871	0.885	0.835	0.836
up	HGSC	0.824	0.934	0.935	0.923	0.849
up	LGSC	0.365	0.78	0.933	0.86	0.73
up	MUC	0.893	0.931	0.94	0.904	0.768
down	CCOC	0.92	0.916	0.913	0.916	0.909
down	ENOC	0.872	0.861	0.882	0.864	0.866
down	HGSC	0.92	0.919	0.912	0.901	0.891
down	LGSC	0.908	0.91	0.931	0.914	0.867
down	MUC	0.925	0.926	0.934	0.914	0.831
smote	CCOC	0.918	0.918	0.923	0.919	0.741
smote	ENOC	0.87	0.867	0.88	0.861	0.666
smote	HGSC	0.941	**0.943**	0.939	0.935	0.714
smote	LGSC	0.827	0.875	0.915	0.895	0.522
smote	MUC	0.931	0.932	0.936	0.924	0.681

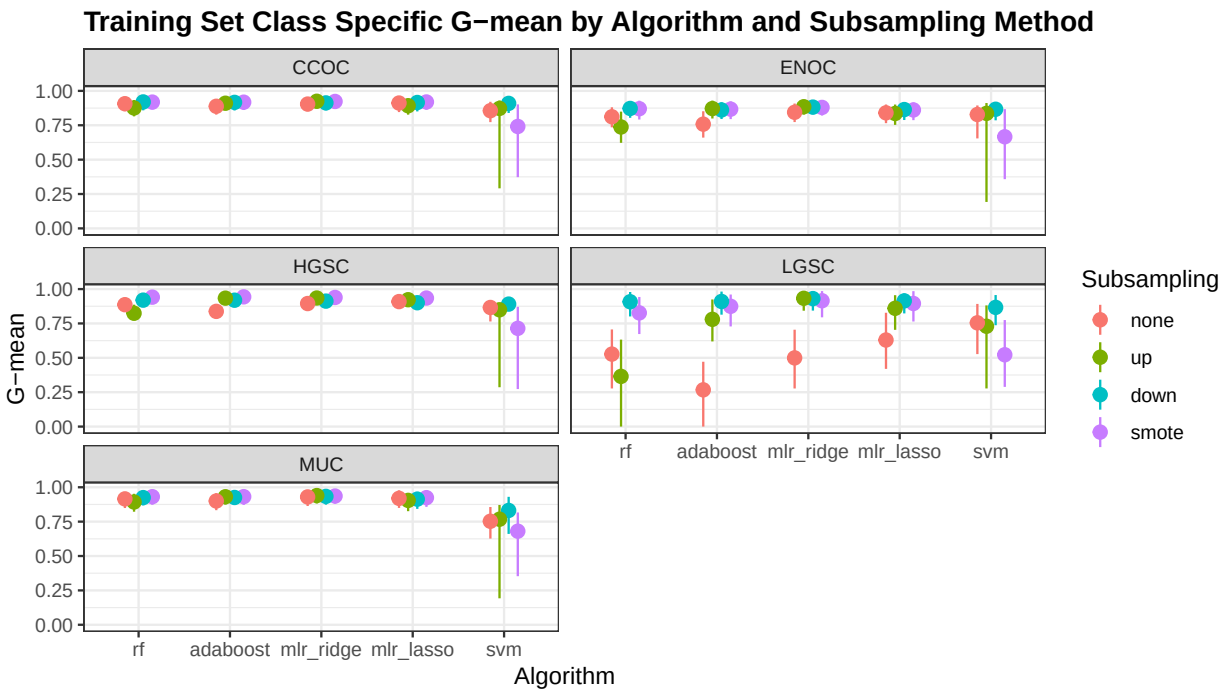


Figure 4.8: Training Set Class-Specific G-mean

Table 4.9: CS1 Set Accuracy by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.828	0.808	0.841	0.831	**0.849**
up	0.847	0.835	0.842	0.824	0.841
down	0.802	0.781	0.788	0.766	0.811
smote	0.846	0.839	0.837	0.823	0.841

4.2 CS1 Set

4.2.1 Accuracy

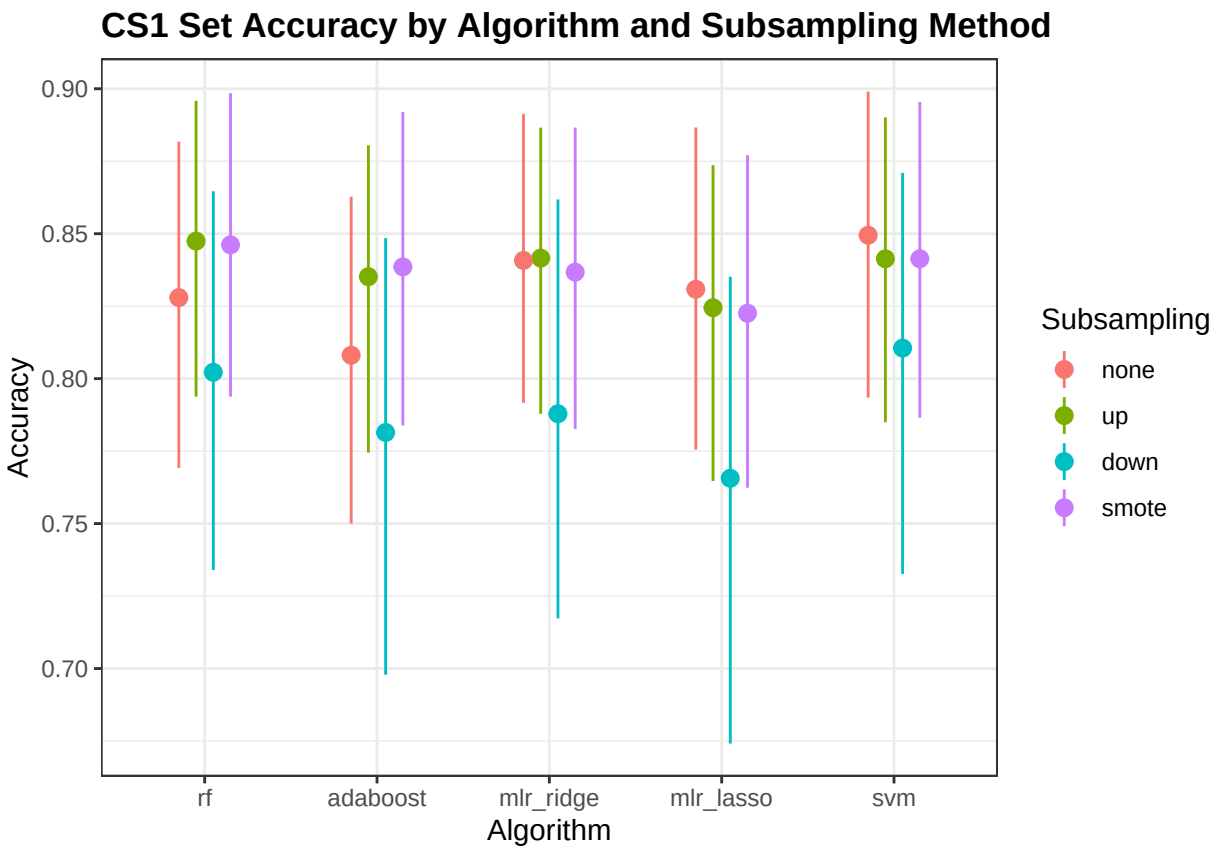


Figure 4.9: CS1 Set Accuracy

Table 4.10: CS1 Set Class-Specific Accuracy by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.942	0.941	0.938	0.937	0.944
none	ENOC	0.891	0.887	0.898	0.897	0.912
none	HGSC	0.902	0.882	0.912	0.904	0.903
none	LGSC	0.956	0.947	0.968	0.957	0.972
none	MUC	0.969	0.967	0.977	0.97	0.969
up	CCOC	0.945	0.941	0.933	0.922	0.937
up	ENOC	0.901	0.892	0.896	0.884	0.904
up	HGSC	0.918	0.906	0.916	0.911	0.899
up	LGSC	0.968	0.965	0.967	0.961	**0.978**
up	MUC	0.971	0.969	0.971	0.977	0.969
down	CCOC	0.939	0.933	0.941	0.926	0.936
down	ENOC	0.881	0.87	0.888	0.873	0.888
down	HGSC	0.888	0.871	0.868	0.856	0.882
down	LGSC	0.941	0.935	0.922	0.92	0.958
down	MUC	0.967	0.96	0.967	0.959	0.96
smote	CCOC	0.944	0.939	0.933	0.931	0.941
smote	ENOC	0.896	0.89	0.894	0.887	0.9
smote	HGSC	0.92	0.913	0.911	0.901	0.901
smote	LGSC	0.968	0.968	0.962	0.957	0.976
smote	MUC	0.97	0.969	0.977	**0.978**	0.969

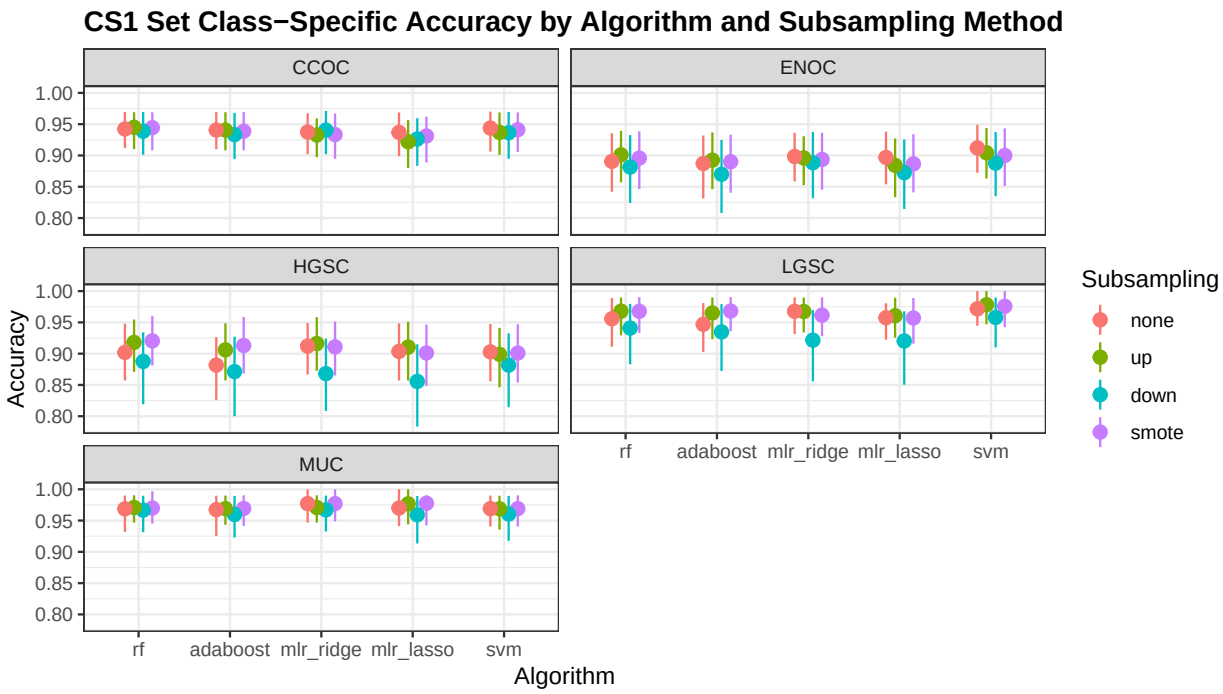


Figure 4.10: CS1 Set Class-Specific Accuracy

Table 4.11: CS1 Set Macro-Averaged F1-Score by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.748	0.718	0.794	0.771	0.805
up	0.792	0.772	**0.806**	0.787	0.793
down	0.76	0.733	0.751	0.723	0.771
smote	0.804	0.797	0.803	0.784	0.797

4.2.2 F1-Score

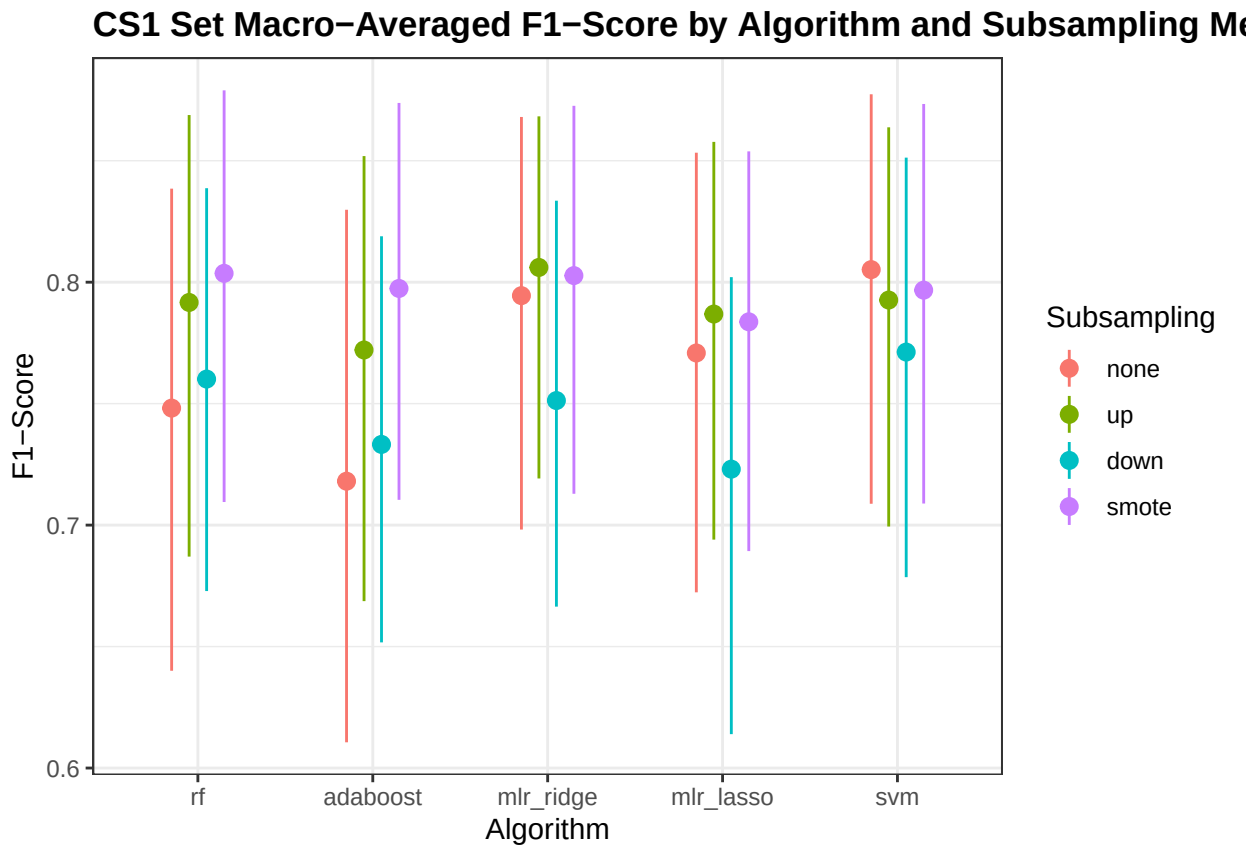


Figure 4.11: CS1 Set F1-Score

Table 4.12: CS1 Set Class-Specific F1-Score by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.829	0.828	0.824	0.813	0.833
none	ENOC	0.764	0.739	0.769	0.769	0.8
none	HGSC	0.9	0.884	0.909	0.9	0.9
none	LGSC	0.545	0.444	0.714	0.625	0.769
none	MUC	0.727	0.667	0.8	0.769	0.727
up	CCOC	0.839	0.828	0.812	0.784	0.813
up	ENOC	0.78	0.765	0.766	0.743	0.782
up	HGSC	**0.915**	0.903	0.909	0.902	0.898
up	LGSC	0.667	0.667	0.769	0.727	0.8
up	MUC	0.769	0.75	0.8	0.778	0.71
down	CCOC	0.828	0.812	0.833	0.8	0.824
down	ENOC	0.743	0.711	0.756	0.723	0.764
down	HGSC	0.871	0.85	0.843	0.83	0.865
down	LGSC	0.667	0.632	0.6	0.571	0.714
down	MUC	0.727	0.714	0.732	0.706	0.727
smote	CCOC	0.839	0.833	0.821	0.811	0.833
smote	ENOC	0.779	0.766	0.769	0.757	0.785
smote	HGSC	**0.915**	0.907	0.901	0.889	0.898
smote	LGSC	0.75	0.75	0.75	0.706	0.769
smote	MUC	0.769	0.766	0.8	0.8	0.727

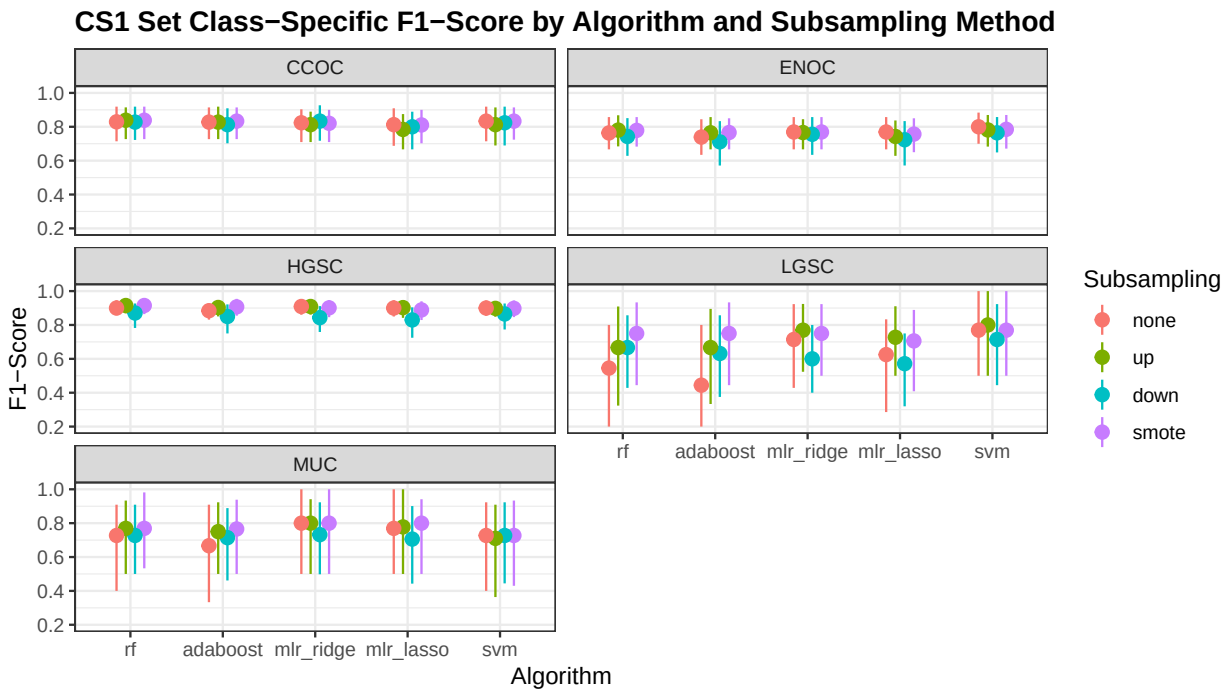


Figure 4.12: CS1 Set Class-Specific F1-Score

Table 4.13: CS1 Set Kappa by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.743	0.706	0.765	0.752	0.776
up	0.775	0.755	0.771	0.747	0.763
down	0.723	0.694	0.707	0.675	0.733
smote	**0.777**	0.766	0.767	0.746	0.768

4.2.3 Kappa

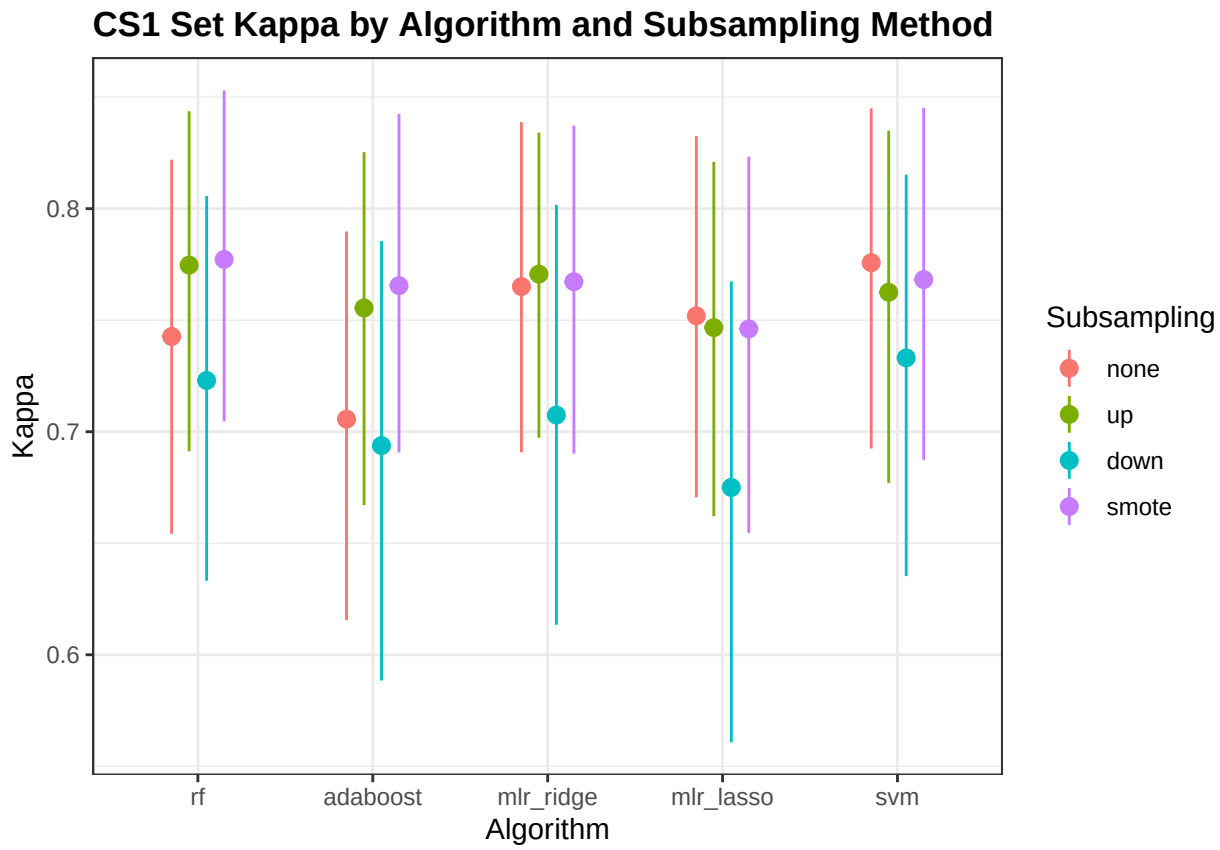


Figure 4.13: CS1 Set Kappa

Table 4.14: CS1 Set Class-Specific Kappa by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.795	0.792	0.784	0.777	0.797
none	ENOC	0.691	0.666	0.704	0.7	0.744
none	HGSC	0.803	0.764	0.824	0.807	0.806
none	LGSC	0.49	0.342	0.692	0.593	0.754
none	MUC	0.709	0.652	0.784	0.753	0.712
up	CCOC	0.804	0.792	0.773	0.734	0.776
up	ENOC	0.713	0.696	0.697	0.664	0.722
up	HGSC	0.836	0.811	0.83	0.82	0.799
up	LGSC	0.652	0.646	0.753	0.711	0.784
up	MUC	0.753	0.73	0.782	0.757	0.678
down	CCOC	0.789	0.774	0.797	0.755	0.784
down	ENOC	0.664	0.624	0.679	0.643	0.691
down	HGSC	0.772	0.738	0.731	0.706	0.76
down	LGSC	0.632	0.594	0.558	0.523	0.691
down	MUC	0.708	0.687	0.712	0.682	0.709
smote	CCOC	0.804	0.794	0.776	0.767	0.796
smote	ENOC	0.709	0.693	0.699	0.683	0.717
smote	HGSC	**0.84**	0.825	0.819	0.799	0.802
smote	LGSC	0.727	0.739	0.728	0.677	0.754
smote	MUC	0.753	0.745	0.789	0.788	0.711

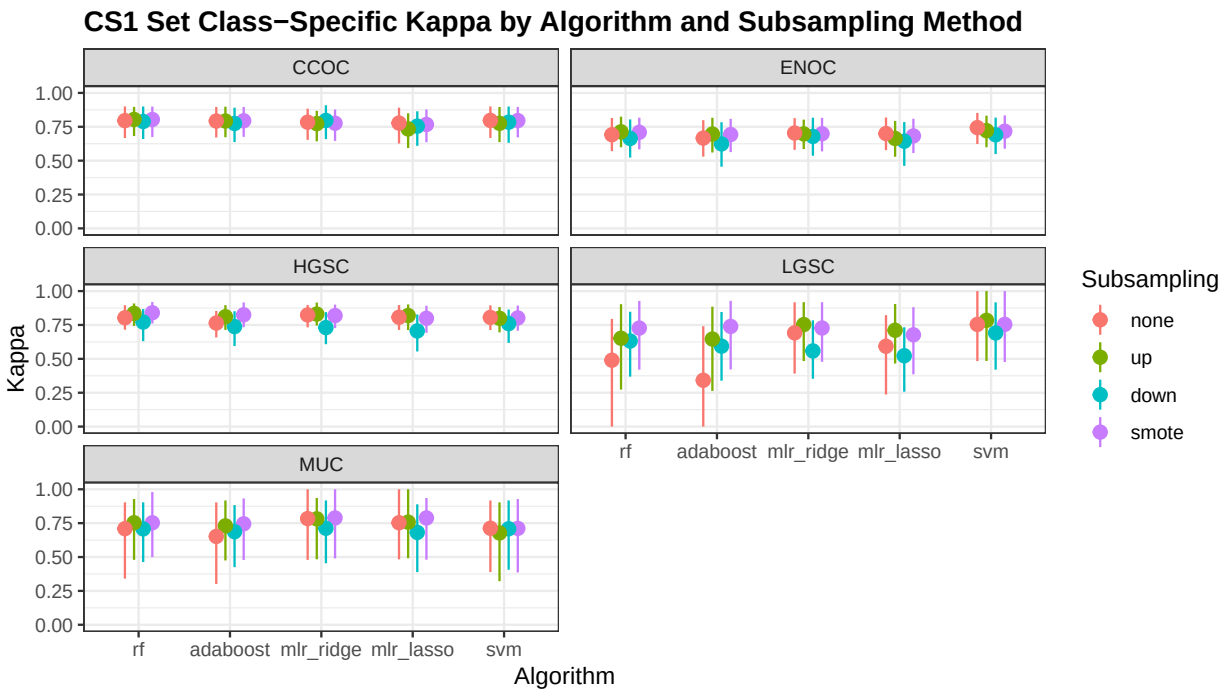


Figure 4.14: CS1 Set Class-Specific Kappa

Table 4.15: CS1 Set G-mean by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.658	0.571	0.75	0.727	0.757
up	0.734	0.716	**0.812**	0.787	0.734
down	0.791	0.774	0.795	0.756	0.793
smote	0.786	0.781	0.81	0.793	0.752

4.2.4 G-mean

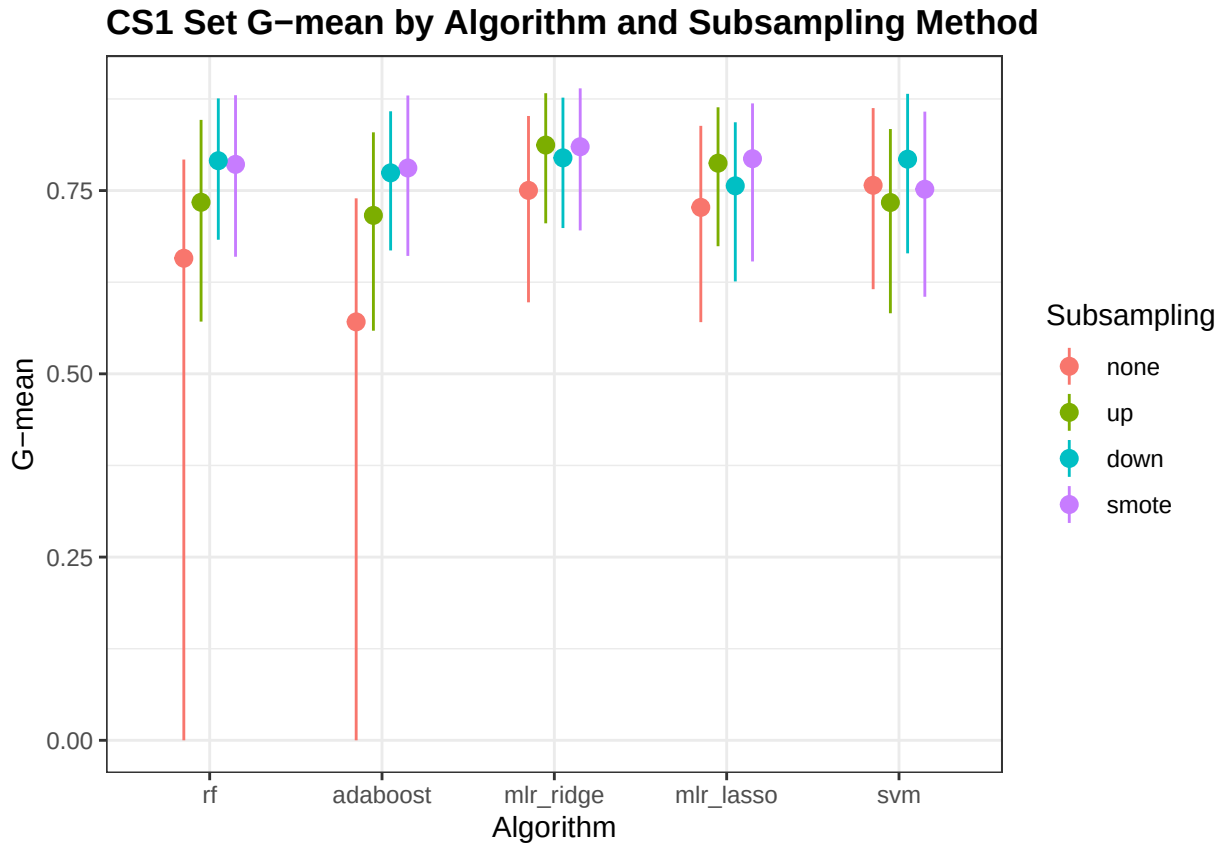


Figure 4.15: CS1 Set G-mean

Table 4.16: CS1 Set Class-Specific G-mean by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.888	0.885	0.889	0.885	0.891
none	ENOC	0.854	0.827	0.848	0.848	0.867
none	HGSC	0.904	0.883	0.913	0.905	0.905
none	LGSC	0.606	0.471	0.775	0.707	0.816
none	MUC	0.775	0.745	0.845	0.84	0.812
up	CCOC	0.893	0.888	0.891	0.87	0.883
up	ENOC	0.864	0.849	0.846	0.833	0.848
up	HGSC	**0.92**	0.908	0.914	0.909	0.902
up	LGSC	0.707	0.707	0.913	0.889	0.816
up	MUC	0.816	0.812	0.878	0.864	0.756
down	CCOC	0.901	0.895	0.905	0.883	0.886
down	ENOC	0.84	0.81	0.849	0.82	0.853
down	HGSC	0.881	0.863	0.855	0.845	0.876
down	LGSC	0.904	0.902	0.91	0.858	0.895
down	MUC	0.856	0.861	0.859	0.852	0.895
smote	CCOC	0.899	0.896	0.895	0.89	0.894
smote	ENOC	0.872	0.856	0.854	0.846	0.868
smote	HGSC	**0.92**	0.913	0.907	0.895	0.902
smote	LGSC	0.835	0.84	0.898	0.863	0.816
smote	MUC	0.856	0.848	0.882	0.877	0.788

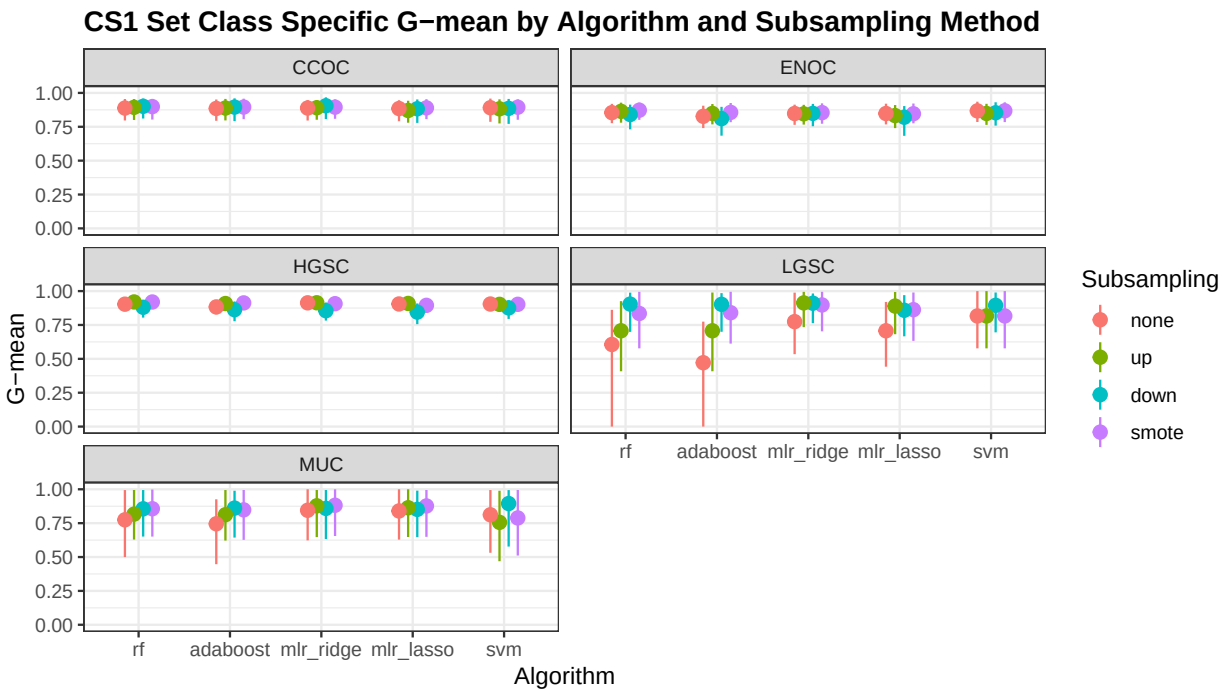


Figure 4.16: CS1 Set Class-Specific G-mean

Table 4.17: CS2 Set Accuracy by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.924	0.91	**0.938**	0.931	0.926
up	0.926	0.931	0.922	0.921	0.926
down	0.859	0.844	0.815	0.817	0.843
smote	0.928	0.925	0.915	0.902	0.922

4.3 CS2 Set

4.3.1 Accuracy

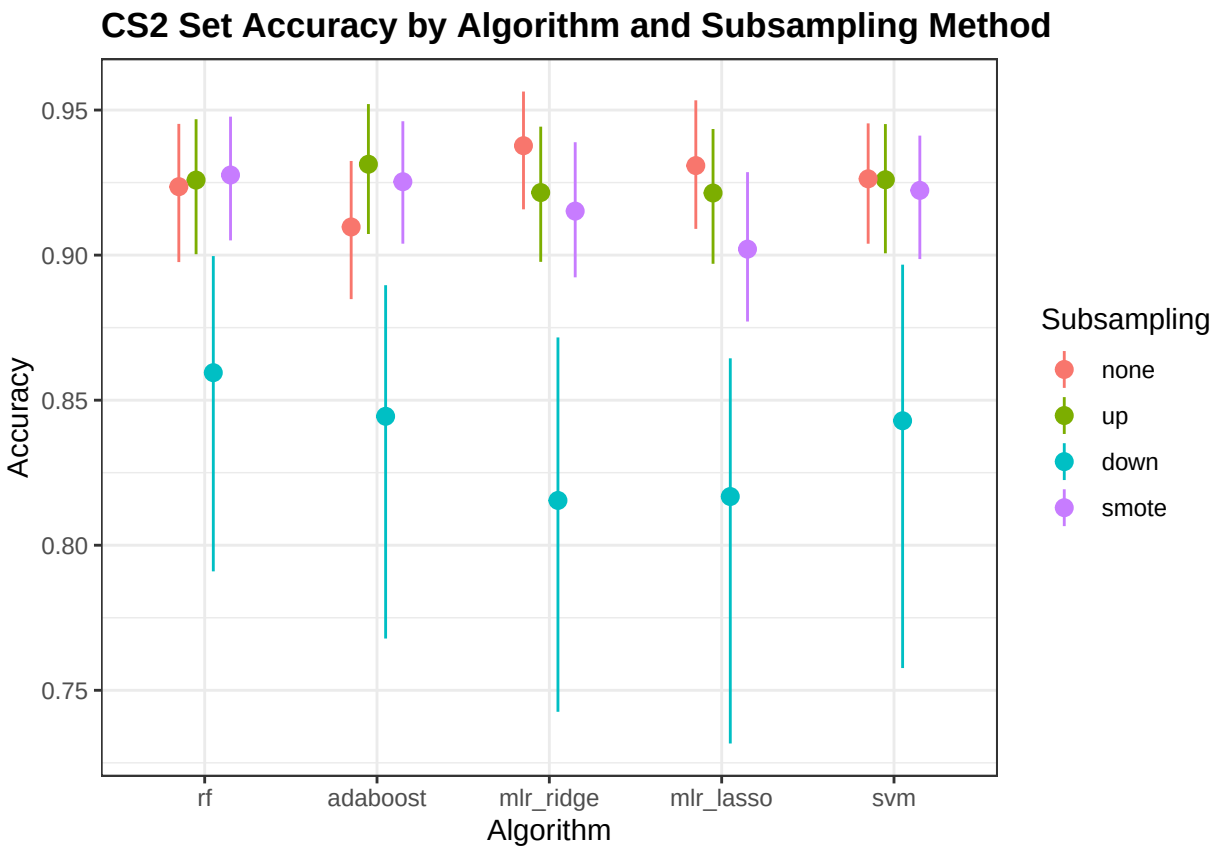


Figure 4.17: CS2 Set Accuracy

Table 4.18: CS2 Set Class-Specific Accuracy by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.984	0.983	**0.987**	0.983	0.981
none	ENOC	0.974	0.967	0.98	0.977	0.981
none	HGSC	0.931	0.913	0.949	0.946	0.936
none	LGSC	0.977	0.976	0.977	0.975	0.977
none	MUC	0.983	0.981	0.983	0.981	0.977
up	CCOC	0.986	0.986	0.986	0.984	0.98
up	ENOC	0.977	0.98	0.969	0.969	0.98
up	HGSC	0.931	0.941	0.938	0.941	0.933
up	LGSC	0.977	0.977	0.972	0.972	0.98
up	MUC	0.981	0.981	0.98	0.979	0.977
down	CCOC	0.98	0.979	0.977	0.97	0.956
down	ENOC	0.96	0.959	0.954	0.943	0.958
down	HGSC	0.879	0.866	0.842	0.844	0.867
down	LGSC	0.948	0.939	0.921	0.922	0.954
down	MUC	0.956	0.951	0.947	0.963	0.961
smote	CCOC	0.984	0.984	0.986	0.981	0.979
smote	ENOC	0.976	0.976	0.966	0.961	0.98
smote	HGSC	0.943	0.941	0.933	0.923	0.934
smote	LGSC	0.979	0.979	0.97	0.964	0.98
smote	MUC	0.974	0.972	0.978	0.976	0.973

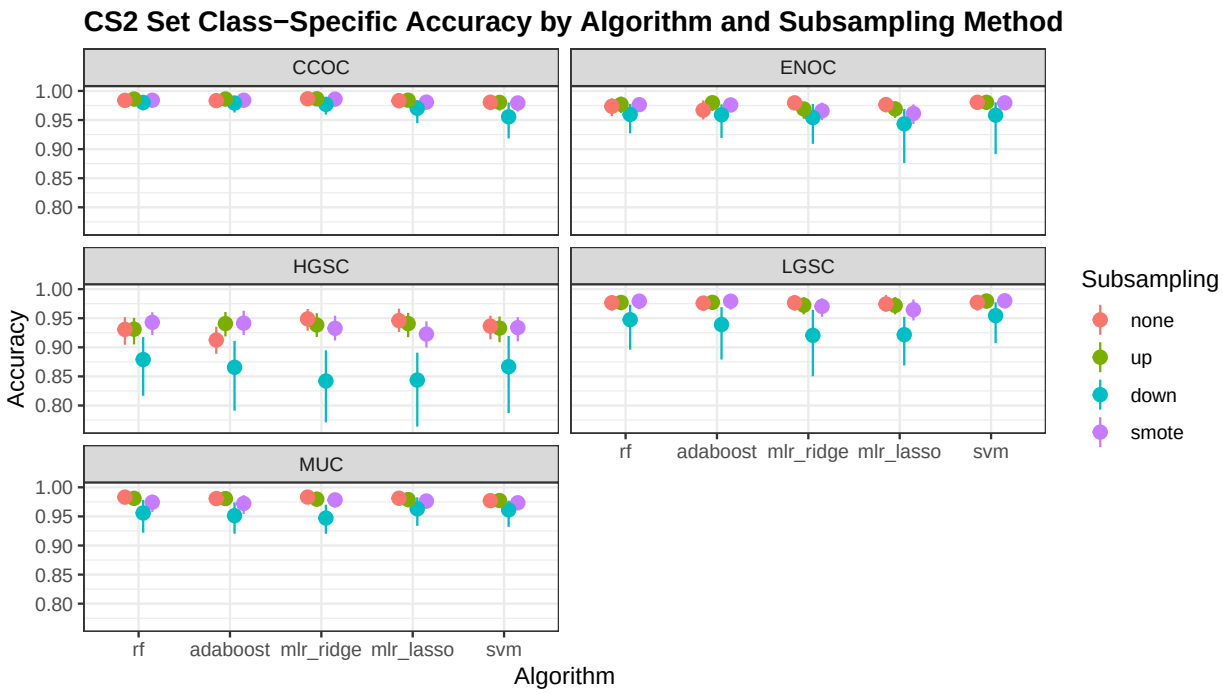


Figure 4.18: CS2 Set Class-Specific Accuracy

Table 4.19: CS2 Set Macro-Averaged F1-Score by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.714	0.751	0.757	0.745	0.766
up	0.722	0.741	**0.784**	0.761	0.755
down	0.703	0.68	0.656	0.645	0.675
smote	0.782	0.775	0.771	0.741	0.758

4.3.2 F1-Score

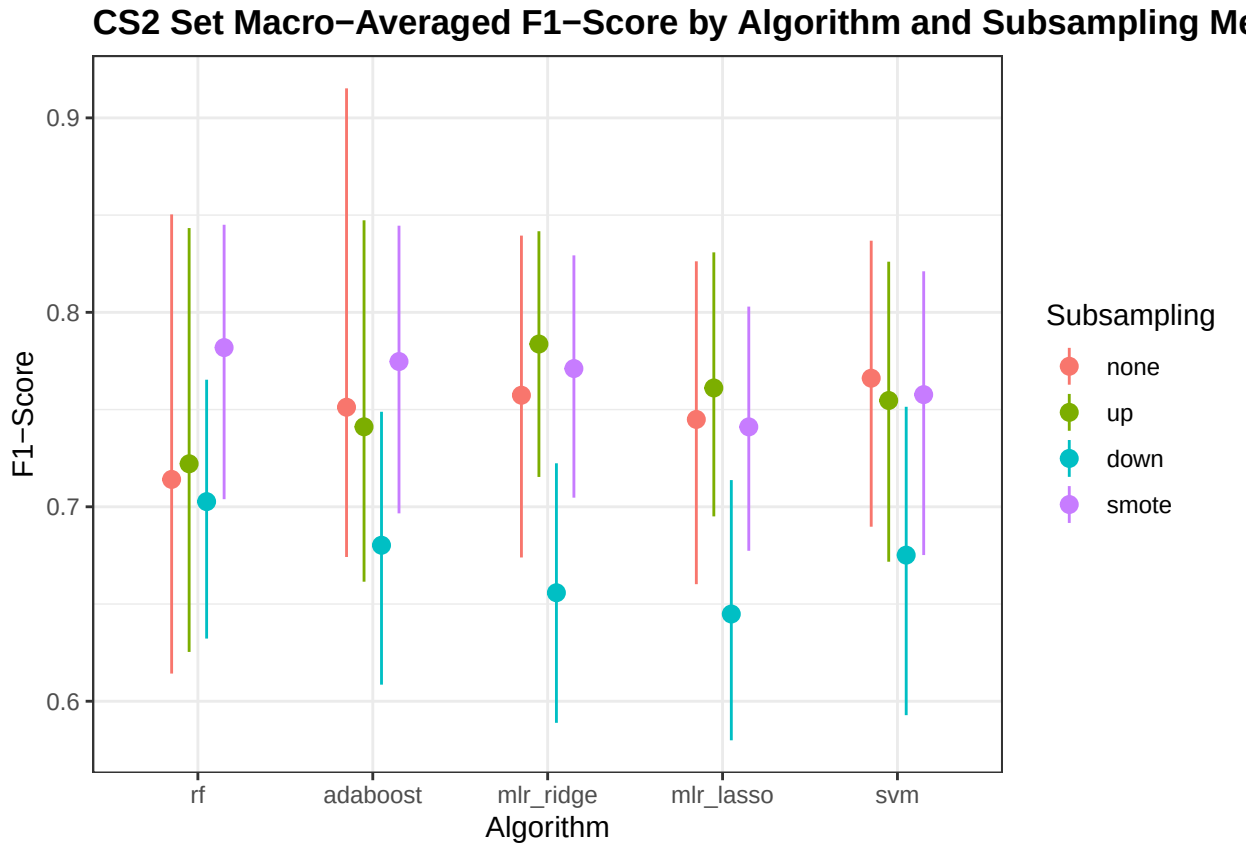


Figure 4.19: CS2 Set F1-Score

Table 4.20: CS2 Set Class-Specific F1-Score by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.889	0.875	0.903	0.884	0.857
none	ENOC	0.471	0.267	0.667	0.632	0.706
none	HGSC	0.958	0.948	**0.968**	0.966	0.961
none	LGSC	0.222	0.286	0.375	0.4	0.5
none	MUC	0.882	0.865	0.885	0.878	0.835
up	CCOC	0.895	0.895	0.909	0.895	0.848
up	ENOC	0.556	0.632	0.609	0.6	0.696
up	HGSC	0.958	0.963	0.96	0.962	0.959
up	LGSC	0.25	0.286	0.583	0.522	0.5
up	MUC	0.87	0.87	0.864	0.857	0.833
down	CCOC	0.87	0.857	0.844	0.809	0.755
down	ENOC	0.556	0.533	0.5	0.444	0.538
down	HGSC	0.919	0.909	0.89	0.893	0.909
down	LGSC	0.432	0.389	0.343	0.333	0.452
down	MUC	0.745	0.724	0.717	0.768	0.75
smote	CCOC	0.895	0.895	0.9	0.872	0.842
smote	ENOC	0.667	0.667	0.588	0.556	0.667
smote	HGSC	0.963	0.963	0.957	0.95	0.959
smote	LGSC	0.556	0.571	0.558	0.5	0.533
smote	MUC	0.837	0.824	0.857	0.842	0.811

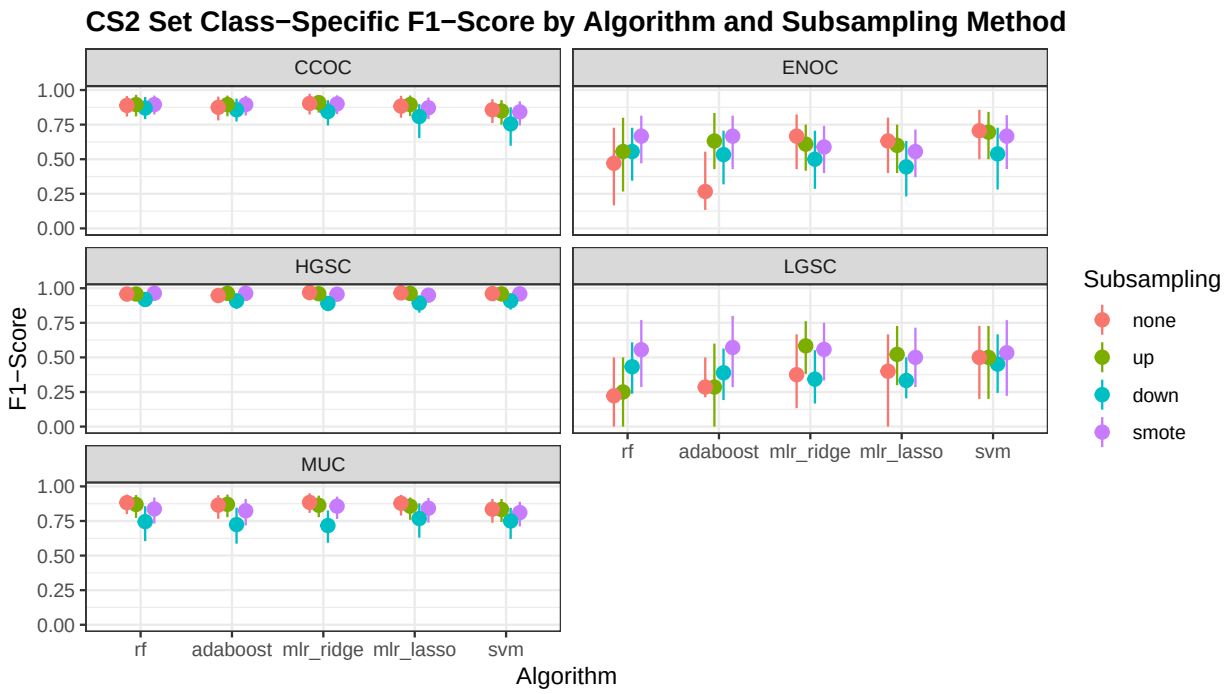


Figure 4.20: CS2 Set Class-Specific F1-Score

Table 4.21: CS2 Set Kappa by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.761	0.703	**0.811**	0.799	0.779
up	0.764	0.794	0.798	0.789	0.767
down	0.676	0.646	0.605	0.602	0.641
smote	0.802	0.797	0.781	0.75	0.766

4.3.3 Kappa

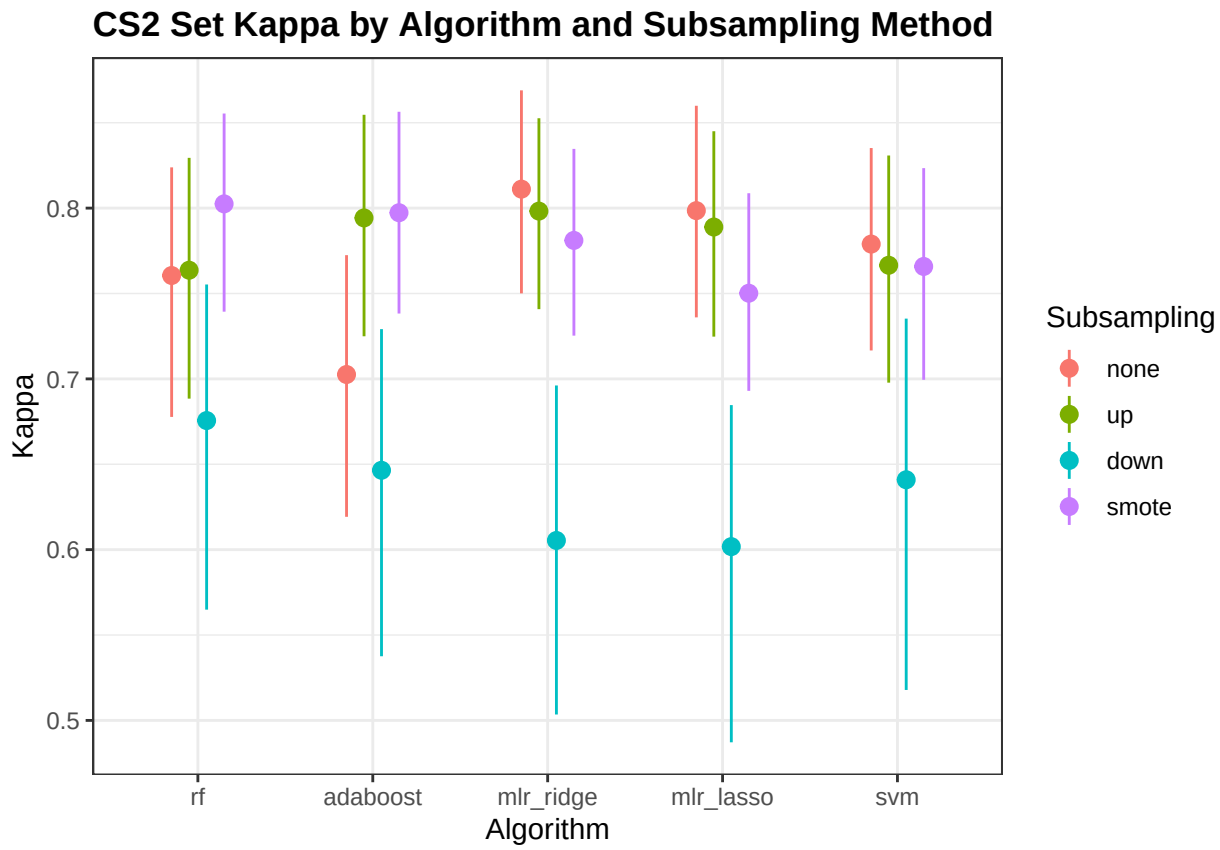


Figure 4.21: CS2 Set Kappa

Table 4.22: CS2 Set Class-Specific Kappa by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.88	0.868	0.896	0.874	0.848
none	ENOC	0.459	0.191	0.655	0.62	0.697
none	HGSC	0.767	0.692	0.834	0.828	0.794
none	LGSC	0.148	0	0.357	0.39	0.486
none	MUC	0.873	0.856	0.875	0.869	0.823
up	CCOC	0.888	0.888	**0.902**	0.885	0.839
up	ENOC	0.544	0.622	0.594	0.586	0.684
up	HGSC	0.764	0.806	0.823	0.823	0.775
up	LGSC	0.214	0.264	0.568	0.506	0.486
up	MUC	0.86	0.859	0.853	0.842	0.823
down	CCOC	0.859	0.847	0.832	0.792	0.732
down	ENOC	0.533	0.512	0.484	0.414	0.516
down	HGSC	0.688	0.659	0.617	0.617	0.656
down	LGSC	0.41	0.364	0.314	0.306	0.429
down	MUC	0.722	0.696	0.687	0.747	0.729
smote	CCOC	0.888	0.887	0.893	0.863	0.832
smote	ENOC	0.655	0.651	0.568	0.54	0.655
smote	HGSC	0.829	0.825	0.81	0.781	0.786
smote	LGSC	0.542	0.561	0.544	0.484	0.523
smote	MUC	0.824	0.808	0.845	0.831	0.797

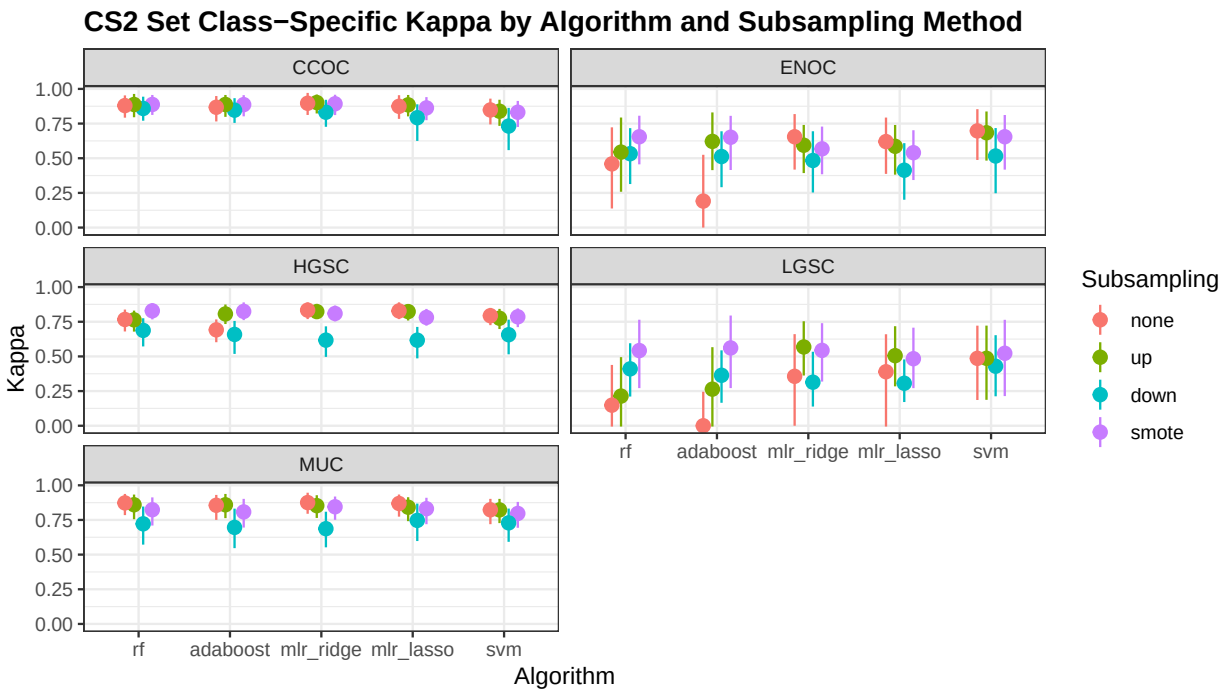


Figure 4.22: CS2 Set Class-Specific Kappa

Table 4.23: CS2 Set G-mean by Algorithm and Subsampling Method

sampling	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	0.363	0	0.649	0.657	0.693
up	0.499	0.576	**0.841**	0.773	0.652
down	0.829	0.811	0.808	0.792	0.802
smote	0.775	0.763	0.835	0.806	0.685

4.3.4 G-mean

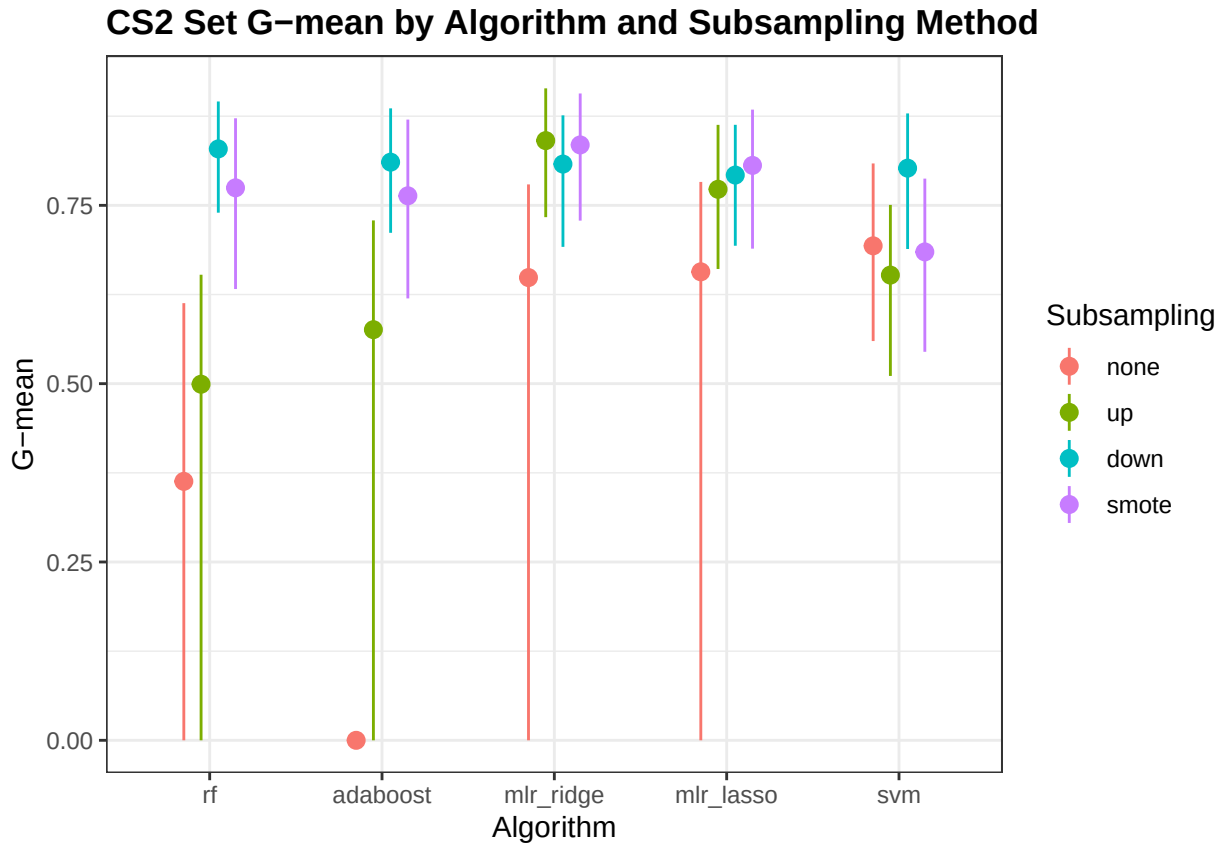


Figure 4.23: CS2 Set G-mean

Table 4.24: CS2 Set Class-Specific G-mean by Algorithm and Subsampling Method

sampling	histotype	rf	adaboost	mlr_ridge	mlr_lasso	svm
none	CCOC	0.913	0.889	0.933	0.928	0.891
none	ENOC	0.576	0.333	0.739	0.742	0.78
none	HGSC	0.83	0.773	0.889	0.894	0.866
none	LGSC	0.301	0	0.535	0.574	0.698
none	MUC	0.921	0.898	0.931	0.93	0.879
up	CCOC	0.911	0.931	**0.96**	0.943	0.866
up	ENOC	0.62	0.707	0.812	0.795	0.755
up	HGSC	0.828	0.871	0.937	0.921	0.84
up	LGSC	0.354	0.408	0.903	0.791	0.629
up	MUC	0.911	0.934	0.941	0.92	0.87
down	CCOC	0.958	0.947	0.928	0.918	0.936
down	ENOC	0.827	0.803	0.807	0.795	0.827
down	HGSC	0.904	0.893	0.881	0.878	0.888
down	LGSC	0.896	0.885	0.887	0.883	0.887
down	MUC	0.916	0.914	0.92	0.901	0.88
smote	CCOC	0.957	0.954	0.957	0.942	0.885
smote	ENOC	0.811	0.79	0.815	0.805	0.739
smote	HGSC	0.922	0.92	0.932	0.921	0.861
smote	LGSC	0.751	0.75	0.891	0.853	0.703
smote	MUC	0.935	0.929	0.934	0.919	0.877

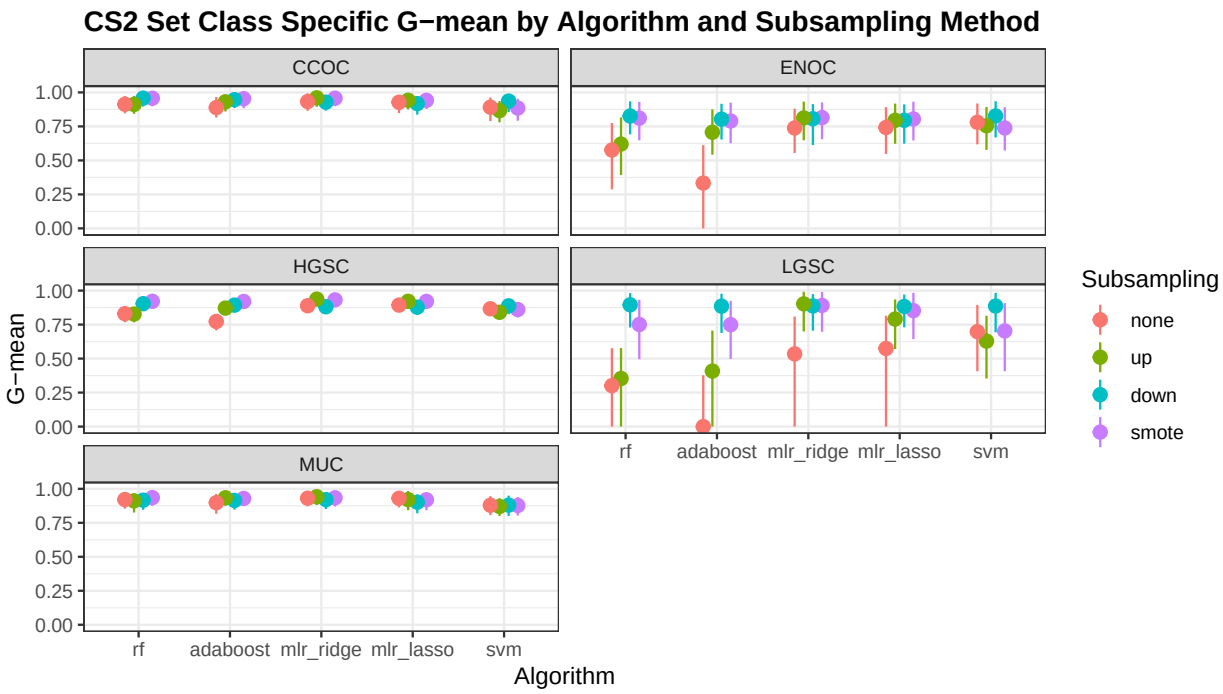


Figure 4.24: CS2 Set Class-Specific G-mean

Table 4.25: SMOTE Kappa by Algorithm and Dataset

dataset	rf	adaboost	mlr_ridge	mlr_lasso	svm
Training	**0.833**	0.823	0.782	0.768	0.561
CS1	0.777	0.766	0.767	0.746	0.768
CS2	0.802	0.797	0.781	0.75	0.766

4.4 SMOTE Kappa Summary

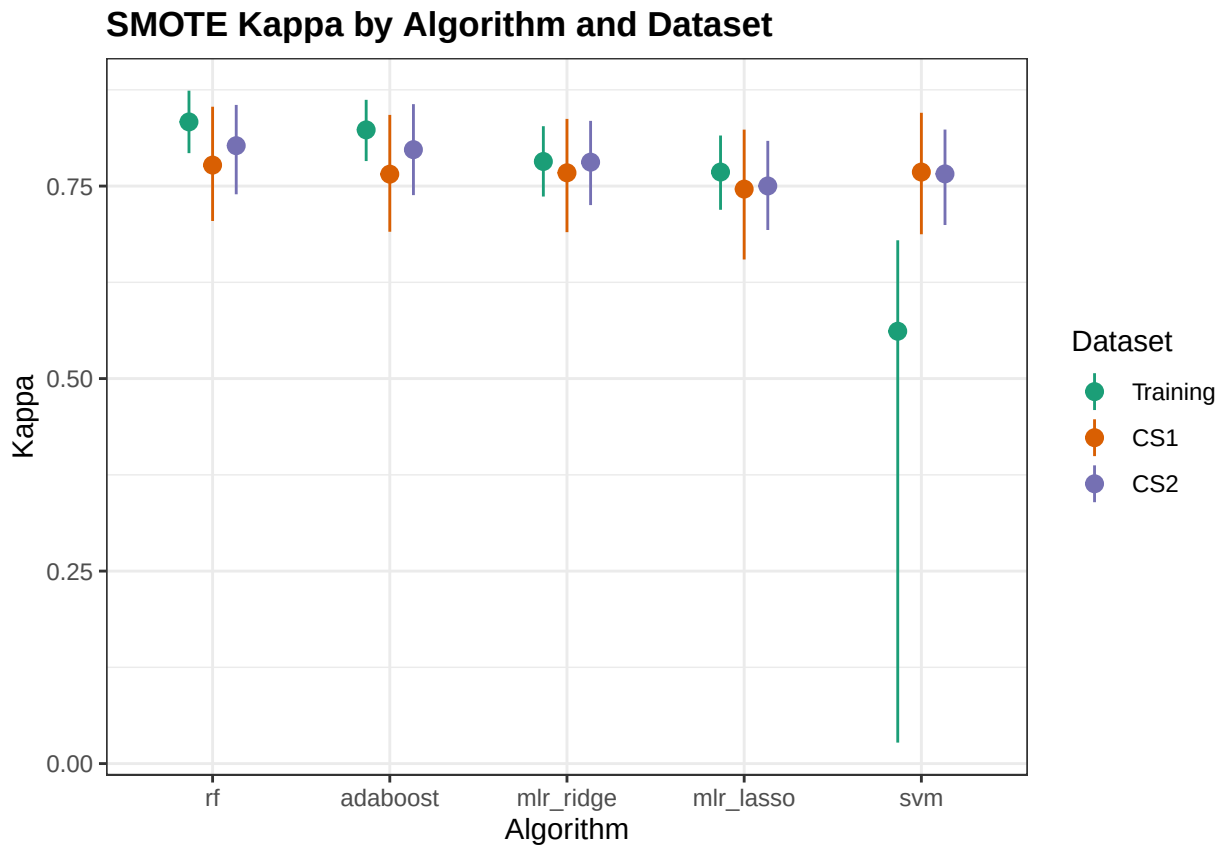


Figure 4.25: SMOTE Kappa by Algorithm and Dataset

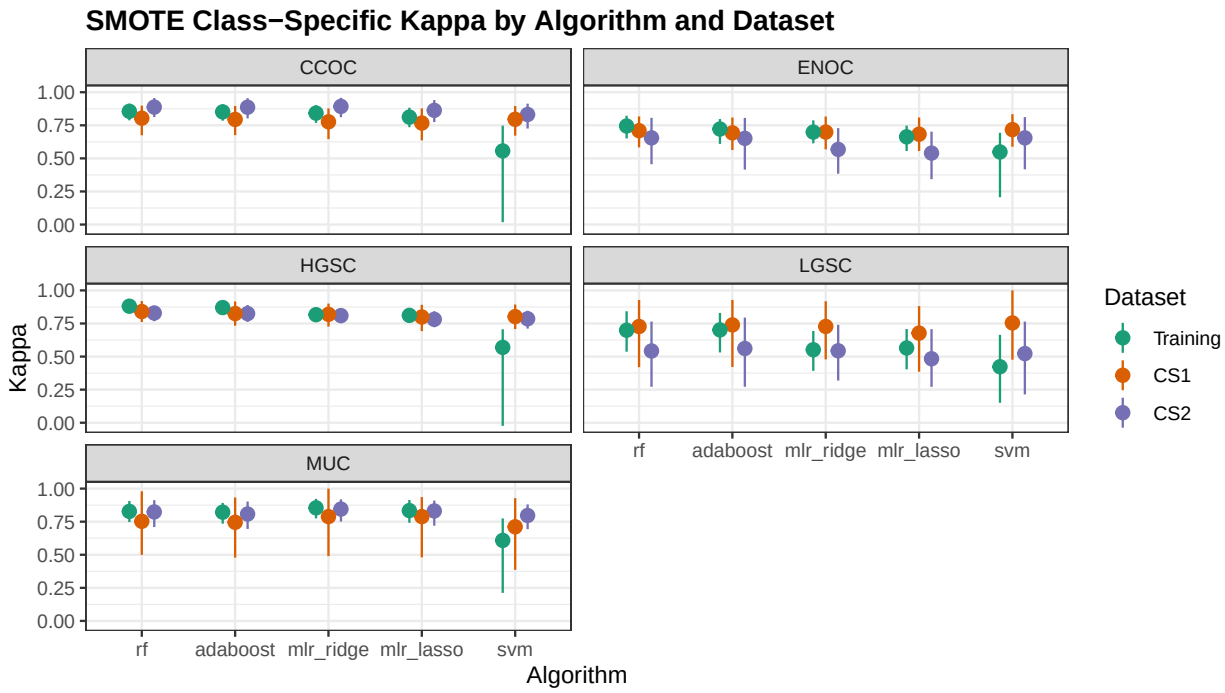


Figure 4.26: SMOTE Class-Specific Kappa by Algorithm and Dataset

4.5 Gene Optimization

4.5.1 Overlap with Other Sets

There are 16 genes out of the 72 common classifier set that overlap with the PrOTYPE classifier: COL11A1, CD74, CD2, TIMP3, LUM, CYTIP, COL3A1, THBS2, TCF7L1, HMGA2, FN1, POSTN, COL1A2, COL5A2, PDZK1IP1, FBN1

There are 13 genes out of the 72 classifier set that overlap with the SPOT signature: HIF1A, CXCL10, DUSP4, SOX17, MITF, CDKN3, BRCA2, CEACAM5, ANXA4, SERPINE1, TCF7L1, CRABP2, DNAJC9.

4.5.2 Optimal Gene Set

There are 28 unique genes from the combined PrOTYPE and SPOT lists that we want to use for the final classifier. We then incrementally add genes from the remaining 44 candidates based on variable importance scores to this list and recalculate performance metrics. The number of genes at which the performance starts to plateau may indicate an optimal gene set for us to carry forward for a particular model.

Variable importance is calculated using either a model-based approach if it is available, or a SHAP-based VI score otherwise (e.g. for SVM). For the sequential and two-step classifiers, we calculate overall VI scores by aggregating the base classifier VI scores using rank aggregation.

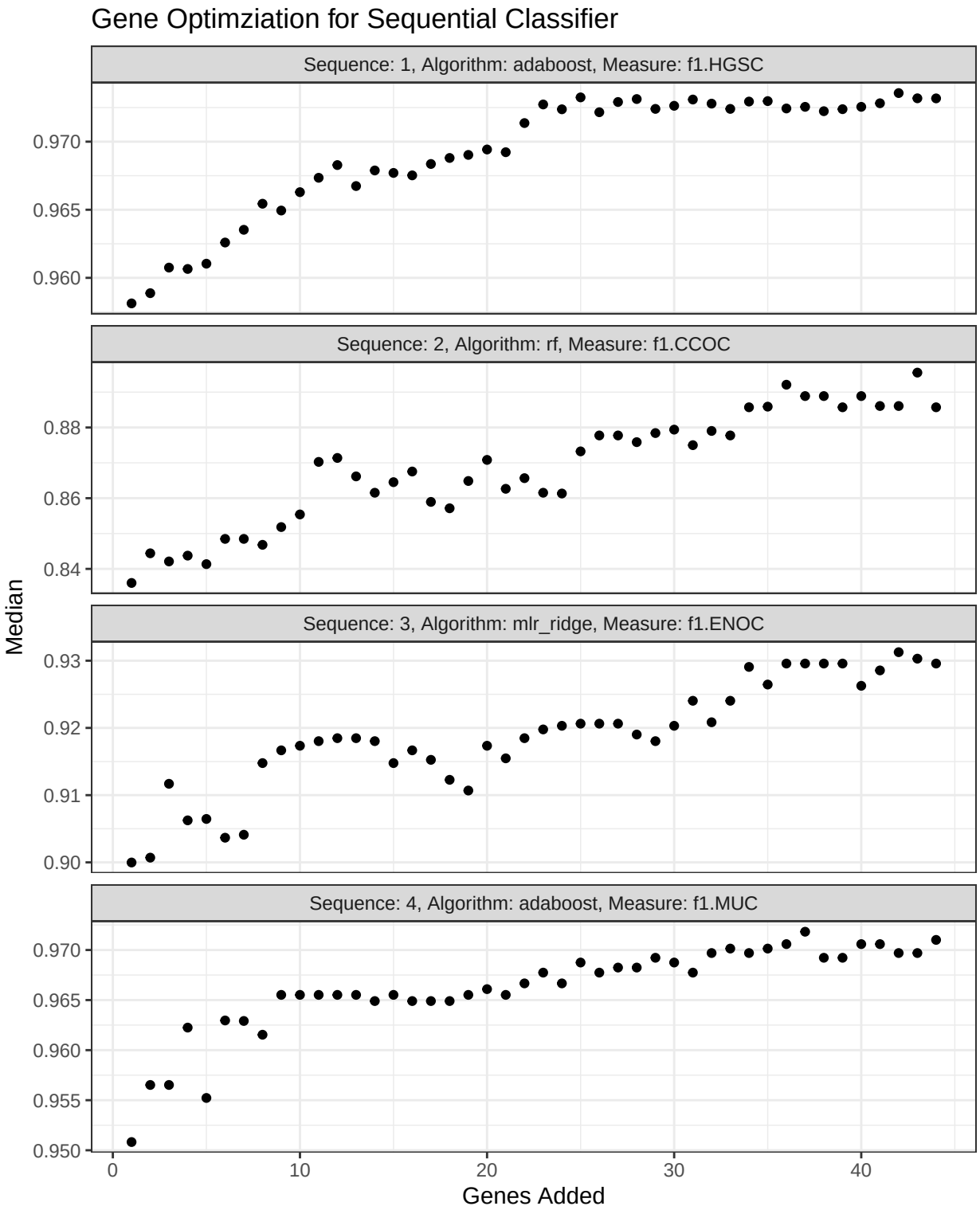


Figure 4.27: Gene Optimization for Sequential Classifier

In the sequential classifier, we use the per-class median F1-scores pertaining to the histotype that had the best performance from each retraining, and sort them on number of genes added. For instance, in sequence 2,

we look at the CCOC F1-scores because CCOC had the best performance from retraining after HGSC was removed.

We can observe that in sequence 3, the F1-score stabilizes at around 0.93 when we reach 34 genes added, hence the optimal number of genes used will be n=28+34=62. The added genes are: SEMA6A, GPR64, KGFLP2, BCL2, ATP5G3, C1orf173, ZBED1, PBX1, FUT3, KLK7, IGFBP1, STC1, MET, CPNE8, C10orf116, MAP1LC3A, EPAS1, SLC3A1, TPX2, TFF1, CAPN2, WT1, CYP4B1, SERPINA5, HNF1B, EGFL6, LGALS4, TSPAN8, BRCA1, LIN28B, DKK4, ADCYAP1R1, TFF3 and MUC5B.

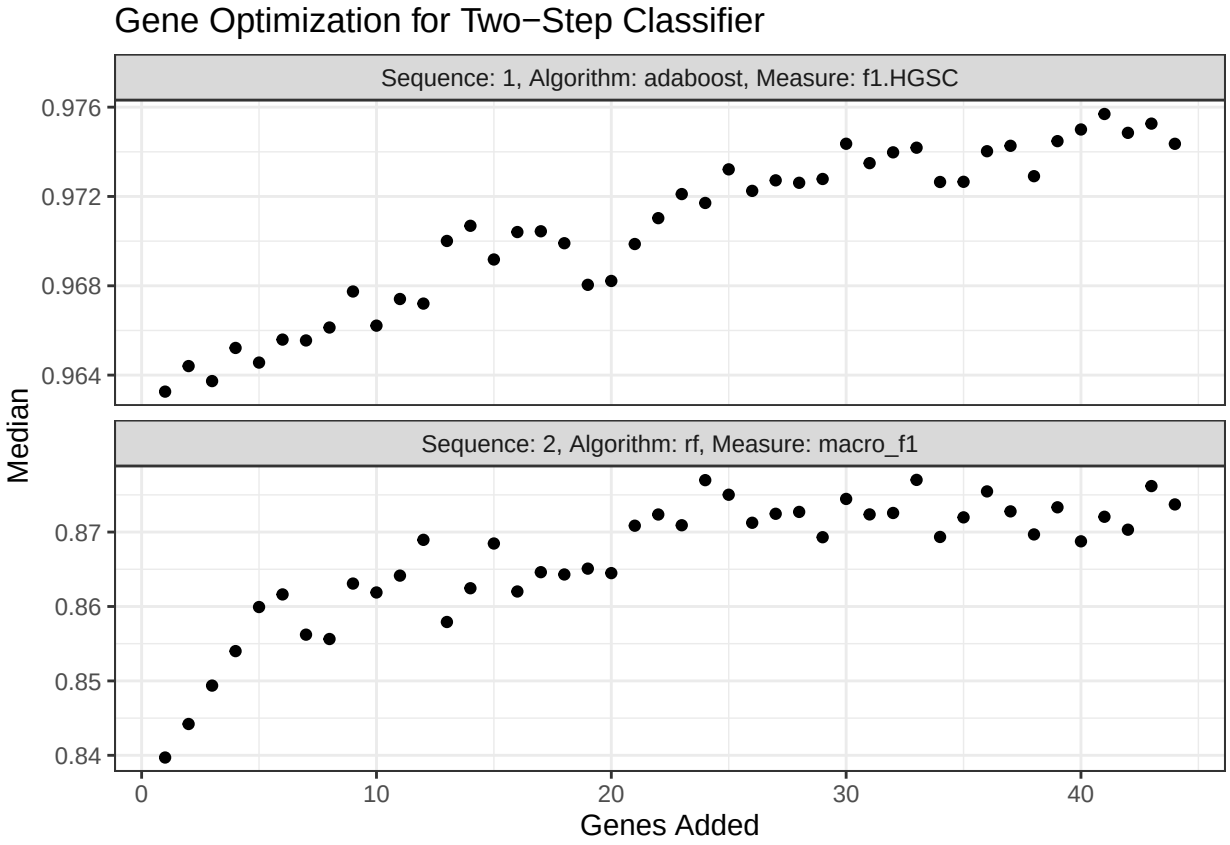


Figure 4.28: Gene Optimization for Two-Step Classifier

In the two-step classifier, we see that in Step 2, the F1-score stabilizes at around 0.88 when we reach 24 added. The optimal number of genes used will be n=28+24=52. The added genes are: PBX1, LGALS4, HNF1B, IGFBP1, TFF3, C10orf116, PAX8, GPR64, FUT3, CYP4B1, DKK4, GAD1, KLK7, EPAS1, CPNE8, BRCA1, ZBED1, IL6, SERPINA5, TPX2, CAPN2, TSPAN8, LIN28B and SLC3A1.

4.6 Rank Aggregation

Show 50 entries Search:

F1-Score Summary by Model and Class

model	CCOC	ENOC	HGSC	LGSC	MUC
All	All	All	All	All	All
seq	0.9	0.931	0.973	0.938	0.971
two_step	0.919	0.837	0.974	0.88	0.897
adaboost-up	0.872	0.769	0.977	0.667	0.857
rf-smote	0.865	0.761	0.975	0.706	0.836
mlr_ridge-none	0.87	0.759	0.97	0.364	0.885
mlr_lasso-none	0.861	0.747	0.971	0.476	0.865
adaboost-smote	0.861	0.738	0.972	0.71	0.831
rf-none	0.871	0.73	0.97	0.417	0.863
mlr_ridge-smote	0.853	0.718	0.958	0.567	0.862
adaboost-none	0.857	0.688	0.961	0.182	0.851
rf-up	0.849	0.677	0.959	0.267	0.846
mlr_lasso-smote	0.825	0.685	0.957	0.579	0.843
mlr_ridge-up	0.844	0.676	0.948	0.545	0.857
svm-none	0.735	0.712	0.958	0.667	0.711
rf-down	0.833	0.63	0.932	0.47	0.8
mlr_lasso-up	0.794	0.648	0.957	0.6	0.821
mlr_ridge-down	0.833	0.644	0.921	0.421	0.838
adaboost-down	0.838	0.628	0.932	0.477	0.776
mlr_lasso-down	0.811	0.615	0.91	0.375	0.794
svm-down	0.632	0.566	0.901	0.514	0.776
svm-up	0.491	0.646	0.905	0.645	0.731
svm-smote	0.58	0.571	0.924	0.429	0.622

Showing 1 to 22 of 22 entries Previous 1 Next

The 22 methods (algorithm-sampling combinations) are ordered in the table by their aggregated ranks using the Genetic Algorithm. We see that the best performing methods involve the 2-stage and sequential algorithms.

4.7 Top 4 Model Summary

4.7.1 Overall Metrics

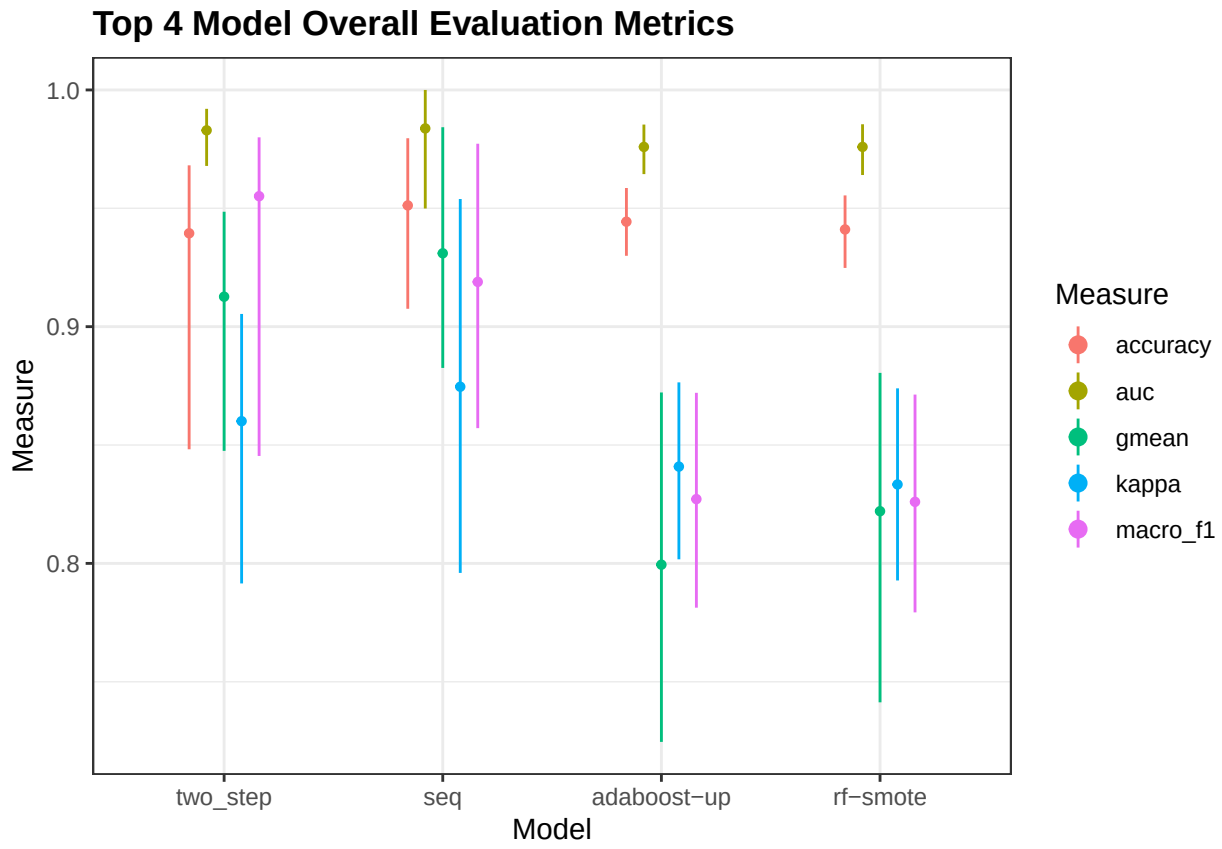


Figure 4.29: Top 4 Model Evaluation Metrics

4.7.2 Per-Class Metrics

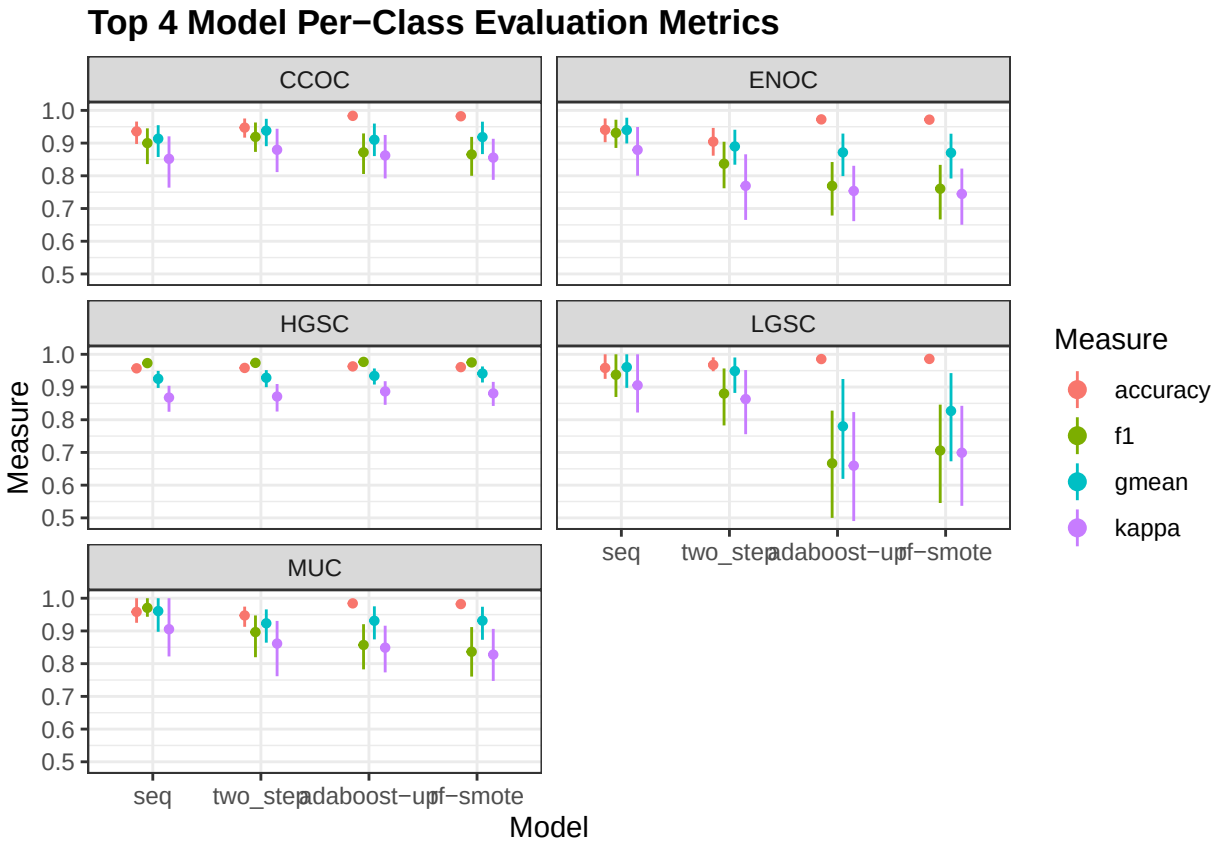


Figure 4.30: Top 4 Model Per-Class Evaluation Metrics

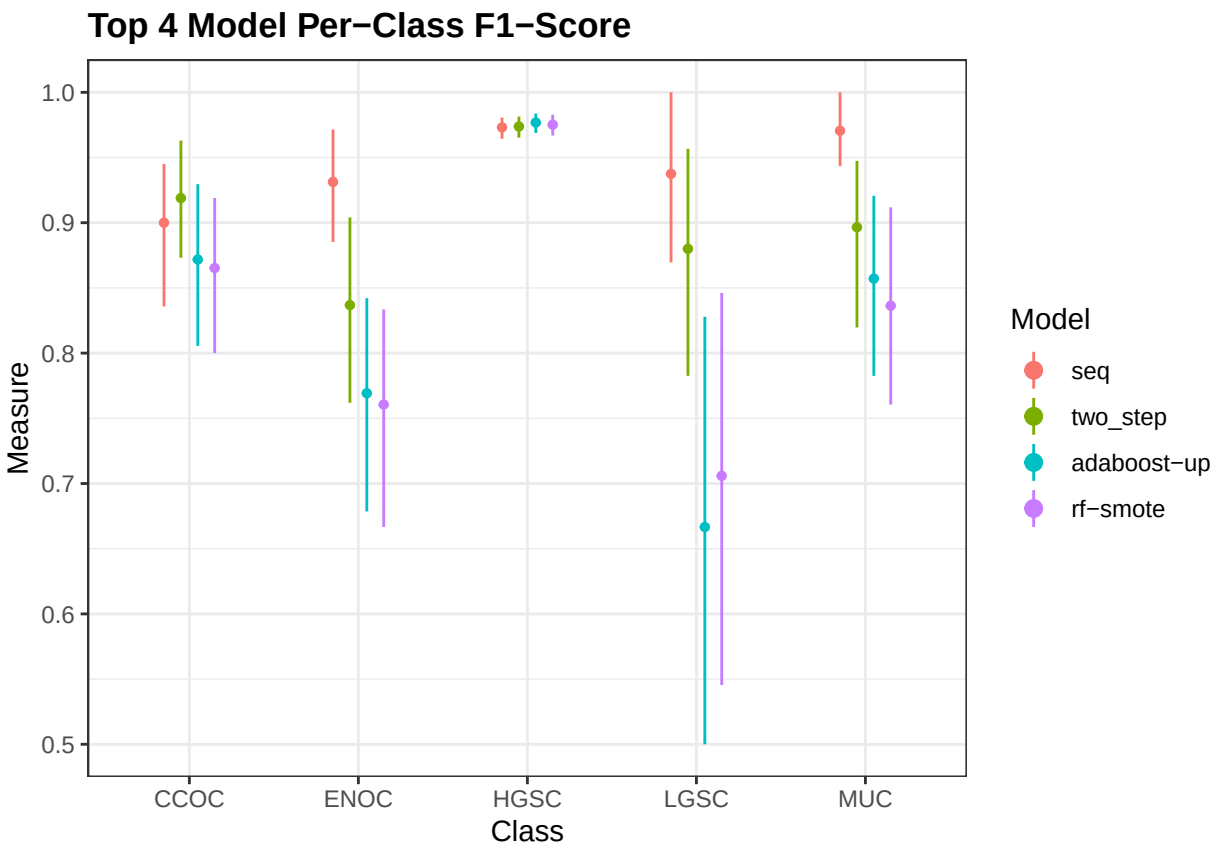


Figure 4.31: Top 4 Model Per-Class F1-Scores

4.8 Test Set Performance

Now we'd like to see how our best methods perform in the confirmation and validation sets. The class-specific F1-scores will be used.

The top 3 methods are:

- **seq**: sequential algorithm with upsampling at every step. The sequence of algorithms used are:
  - HGSC vs. non-HGSC using adaboost
  - CCOC vs. non-CCOC using random forest
  - ENOC vs. non-ENOC using ridge regression
  - MUC vs. LGSC using adaboost
- **two_step**: two-step algorithm with upsampling at both steps. The sequence of algorithms used are:
  - HGSC vs. non-HGSC using adaboost
  - CCOC vs. ENOC vs. MUC vs. LGSC using random forest
- **adaboost-up**: adaboost algorithm with upsampling.

4.8.1 Confirmation Set

In the confirmation set, **sequential_full** and **sequential_optimal** are very similar. Both sequential algorithms have moderate improvement in LGSC and MUC classification. We will select the

Table 4.26: Class-specific F1-scores on Confirmation Set Models

method	HGSC	CCOC	ENOC	LGSC	MUC
two_step_full	0.908	0.886	0.792	0.450	0.679
two_step_optimal	0.906	0.892	0.814	0.486	0.655
sequential_full	0.906	0.922	0.833	0.692	0.852
sequential_optimal	0.906	0.914	0.840	0.692	0.852

Table 4.27: Class-specific F1-scores on Validation Set Model

method	HGSC	CCOC	ENOC	LGSC	MUC
sequential_opt	0.931	0.962	0.919	0.909	0.945

sequential_optimal model to test in the validation set.

4.8.2 Validation Set

Per-class F1-scores in the validation set are all above 0.9.