Ovarian Cancer Histotypes: Report of Statistical Findings

Derek Chiu

2021-12-10

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 Set

The training set distributions for CS1 and CS2 are shown below.

3.3 Common Samples

3.4 Histotypes in Classifier Data

Table 3.1: All CodeSet Histotype Groups

Histotype Group	CS1	CS2	CS3
HGSC	126	655	1779
non-HGSC	168	223	639

Table 3.2: All CodeSet Major Reviewed Histotypes

Reviewed Histotype	CS1	CS2	CS3	CS1 %	CS2 %	CS3 %
CCOC	48	61	181	17.5	7.3	7.7
ENOC	60	34	268	21.8	4.1	11.4
HGSC	126	655	1779	45.8	78.6	75.8
LGSC	21	21	42	7.6	2.5	1.8
MUC	20	62	77	7.3	7.4	3.3

Table 3.3: All CodeSet Reviewed Histotypes

Reviewed Histotype	CS1	CS2	CS3
CARCINOMA-NOS	0	1	23
CCOC	48	61	181
CTRL	0	12	0
ENOC	60	34	268
HGSC	126	655	1779
LGSC	21	21	42
MBOT	0	20	3
MIXED (ENOC/CCOC)	0	0	1
MIXED (ENOC/LGSC)	0	0	1
MIXED (HGSC/CCOC)	0	0	1
MMMT	0	0	30
MUC	20	62	77
Other (use when 6, 7, or 9 is not distinguished) or unknown if epithelial	0	0	1
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	126
CS1	LGSC	21
CS1	MUC	20
CS1	SBOT	19

Table 3.5: CS2 Histotypes

CodeSet	Reviewed Histotype	n
CS2	CARCINOMA-NOS	1
CS2	CCOC	61
CS2	CTRL	12
CS2	ENOC	34
CS2	HGSC	655
CS2	LGSC	21
CS2	MBOT	20
CS2	MUC	62
CS2	SBOT	12

Table 3.6: CS3 Histotypes

CodeSet	Reviewed Histotype	n
CS3	CARCINOMA-NOS	23
CS3	CCOC	181
CS3	ENOC	268
CS3	HGSC	1779
CS3	LGSC	42
CS3	MBOT	3
CS3	MIXED (ENOC/CCOC)	1
CS3	MIXED (ENOC/LGSC)	1
CS3	MIXED (HGSC/CCOC)	1
CS3	MMMT	30
CS3	MUC	77
CS3	Other (use when 6, 7, or 9 is not distinguished) or unknown if epithelial	1
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	59	62	95
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	57	60	73
LGSC	7	5	4
MUC	7	5	5

Table 3.11: Distinct Common Samples Histotype Distribution

revHist	CS1	CS2	CS3
CCOC	3	3	3
ENOC	3	3	3
HGSC	55	55	55
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	27
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	46	46	46
NA	26	26	26

Table 3.16: CS3/CS4/CS5 Pools Distribution

Pool	CS3	CS4	CS5
Pool1	12	5	4
Pool2	5	5	4
Pool3	5	5	4
Pool4	NA	2	1
Pool5	NA	2	1
Pool6	NA	2	0
Pool7	NA	2	1
Pool8	NA	2	1
Pool9	NA	2	1
Pool10	NA	2	1
Pool11	NA	2	1

Table 3.17: Full Training Set Histotype Distribution by CodeSet

Variable	Levels	CS1	CS2	CS3	Total
Histotype	HGSC	122 (49%)	629 (80%)	476 (94%)	1227 (79%)
	CCOC	44 (18%)	54 (7%)	8 (2%)	106 (7%)
	ENOC	55 (22%)	28 (4%)	8 (2%)	91 (6%)
	MUC	16 (6%)	59 (7%)	9 (2%)	84 (5%)
	LGSC	14 (6%)	19 (2%)	6 (1%)	39 (3%)
Total	N (%)	251 (16%)	789 (51%)	507 (33%)	1547 (100%)

Table 3.18: Histotype Distribution by CodeSet/Datasets

Variable	Levels	CS1 All	CS2 All	Confirmation	Validation
Histotype	HGSC	125 (47%)	654 (79%)	423 (66%)	781 (74%)
	CCOC	47 (18%)	60 (7%)	75 (12%)	86 (8%)
	ENOC	58 (22%)	32 (4%)	106 (16%)	140 (13%)
	MUC	19 (7%)	61 (7%)	27 (4%)	34 (3%)
	LGSC	19 (7%)	20 (2%)	13 (2%)	20 (2%)
Total	N (%)	268 (10%)	827 (30%)	644 (23%)	1061 (38%)

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

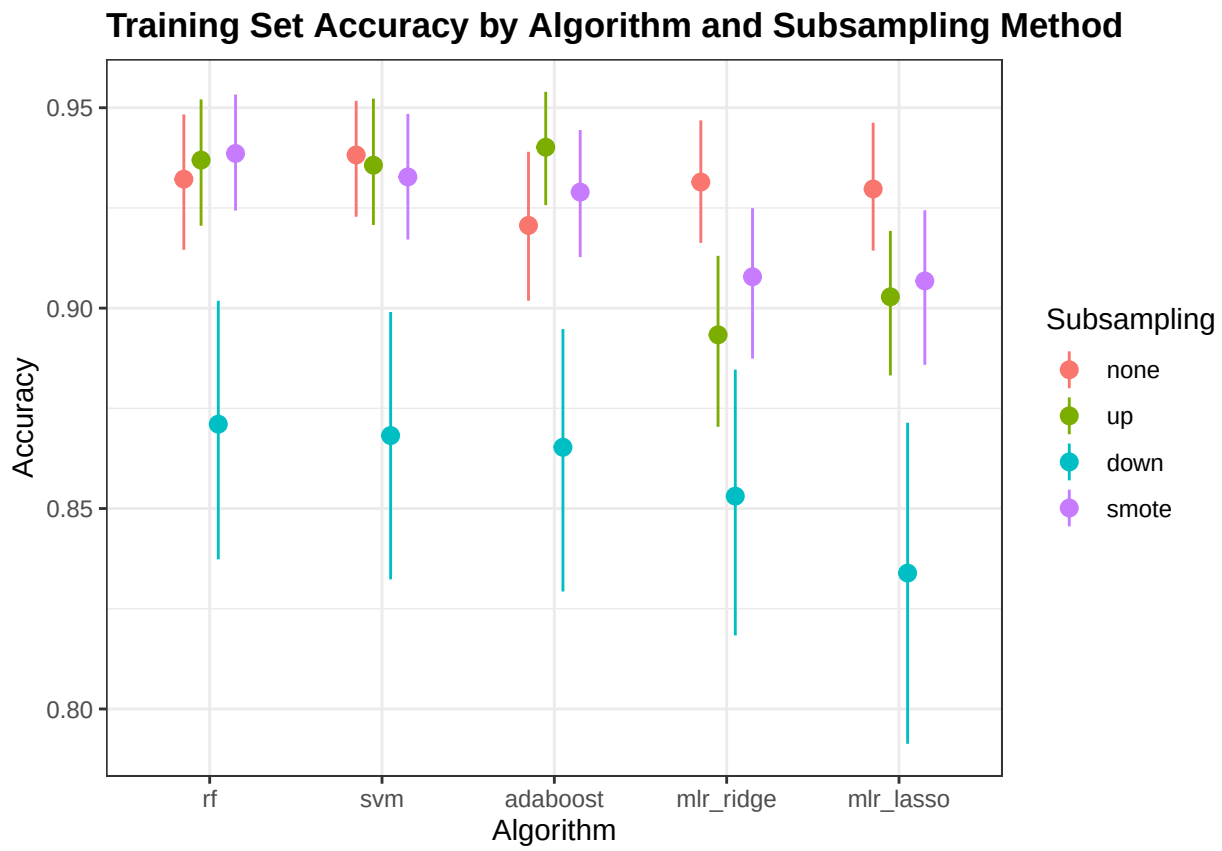

Figure 4.1: Training Set Accuracy

Table 4.1: Training Set Accuracy by Algorithm and Subsampling Method

sampling	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	0.932	0.938	0.921	0.931	0.93
up	0.937	0.936	**0.94**	0.893	0.903
down	0.871	0.868	0.865	0.853	0.834
smote	0.939	0.933	0.929	0.908	0.907

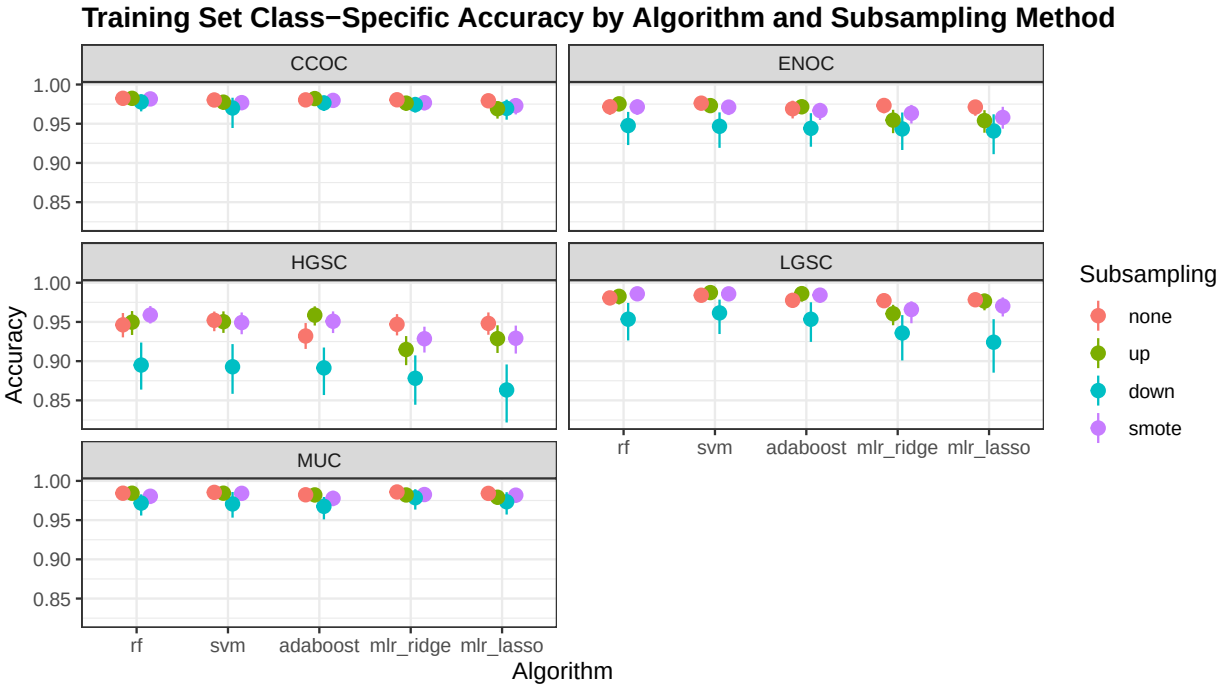

Figure 4.2: Training Set Class-Specific Accuracy

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

sampling	histotype	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	CCOC	0.982	0.98	0.98	0.981	0.979
none	ENOC	0.972	0.976	0.969	0.973	0.971
none	HGSC	0.946	0.952	0.932	0.947	0.948
none	LGSC	0.981	0.984	0.978	0.977	0.978
none	MUC	0.984	0.985	0.982	0.986	0.984
up	CCOC	0.982	0.978	0.982	0.976	0.969
up	ENOC	0.975	0.973	0.972	0.955	0.954
up	HGSC	0.95	0.95	0.959	0.915	0.929
up	LGSC	0.983	**0.987**	0.986	0.96	0.977
up	MUC	0.984	0.984	0.982	0.982	0.979
down	CCOC	0.978	0.97	0.976	0.974	0.97
down	ENOC	0.948	0.947	0.944	0.943	0.941
down	HGSC	0.895	0.893	0.891	0.878	0.863
down	LGSC	0.953	0.962	0.953	0.936	0.924
down	MUC	0.972	0.971	0.968	0.979	0.973
smote	CCOC	0.982	0.977	0.98	0.977	0.973
smote	ENOC	0.971	0.971	0.967	0.963	0.958
smote	HGSC	0.959	0.949	0.951	0.929	0.929
smote	LGSC	0.986	0.986	0.984	0.966	0.97
smote	MUC	0.98	0.984	0.978	0.983	0.982

4.1.2 F1-Score

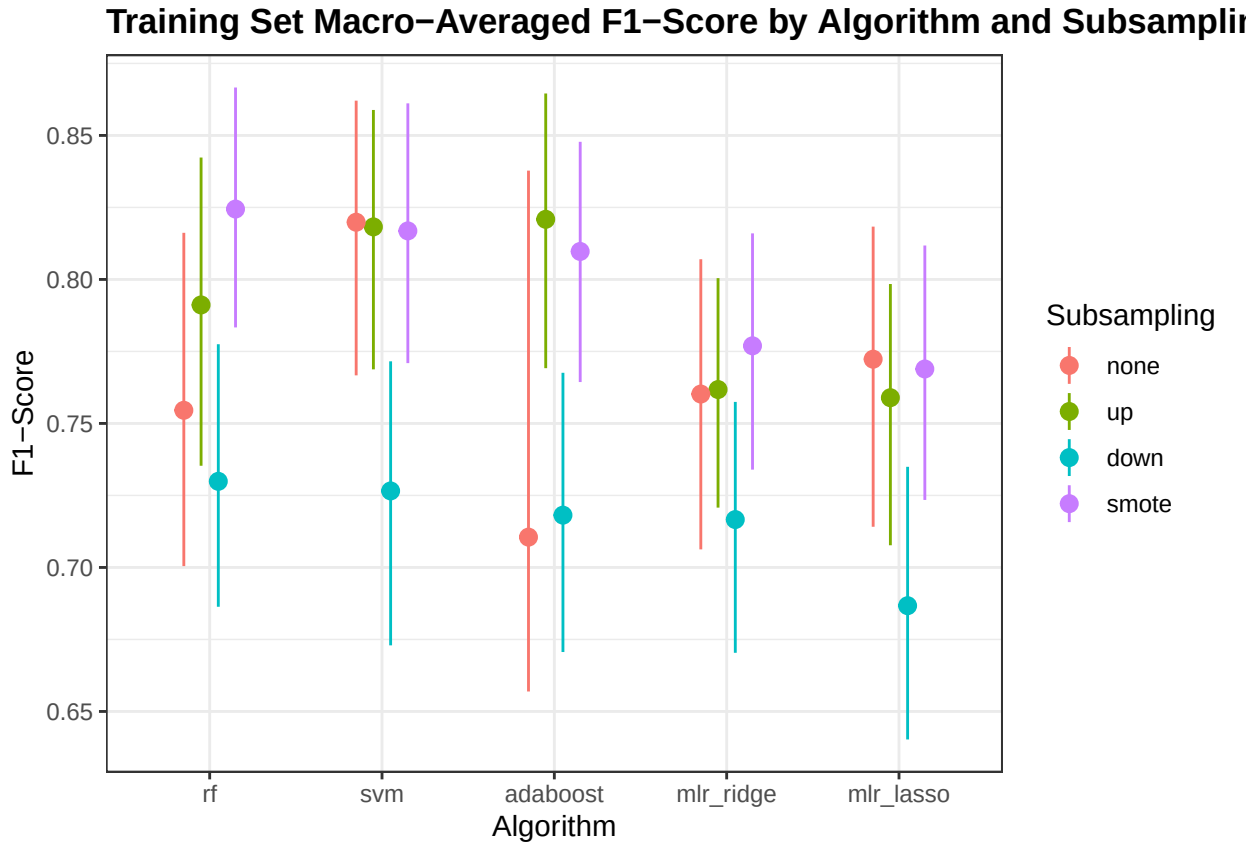

Figure 4.3: Training Set F1-Score

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

sampling	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	0.755	0.82	0.711	0.76	0.772
up	0.791	0.818	0.821	0.762	0.759
down	0.73	0.727	0.718	0.717	0.687
smote	**0.824**	0.817	0.81	0.777	0.769

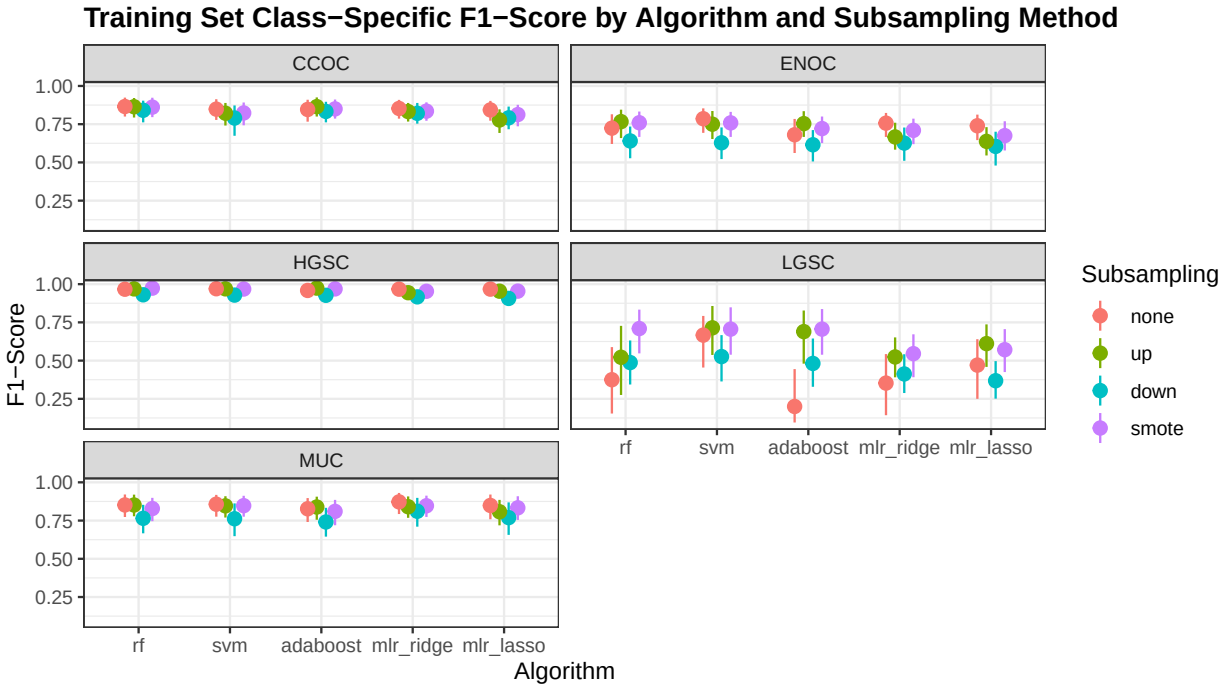

Figure 4.4: Training Set Class-Specific F1-Score

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

sampling	histotype	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	CCOC	0.866	0.848	0.846	0.853	0.843
none	ENOC	0.724	0.786	0.681	0.757	0.74
none	HGSC	0.967	0.97	0.959	0.967	0.968
none	LGSC	0.375	0.667	0.2	0.353	0.471
none	MUC	0.852	0.857	0.827	0.872	0.849
up	CCOC	0.864	0.822	0.865	0.833	0.778
up	ENOC	0.767	0.75	0.754	0.667	0.638
up	HGSC	0.969	0.969	**0.974**	0.944	0.954
up	LGSC	0.522	0.714	0.69	0.524	0.611
up	MUC	0.852	0.846	0.839	0.841	0.808
down	CCOC	0.841	0.79	0.833	0.821	0.792
down	ENOC	0.641	0.629	0.615	0.627	0.605
down	HGSC	0.93	0.928	0.927	0.917	0.907
down	LGSC	0.487	0.526	0.481	0.413	0.368
down	MUC	0.765	0.762	0.741	0.811	0.769
smote	CCOC	0.862	0.824	0.85	0.835	0.813
smote	ENOC	0.759	0.759	0.721	0.71	0.675
smote	HGSC	**0.974**	0.968	0.969	0.953	0.954
smote	LGSC	0.71	0.706	0.706	0.545	0.571
smote	MUC	0.829	0.847	0.81	0.847	0.833

4.1.3 Kappa

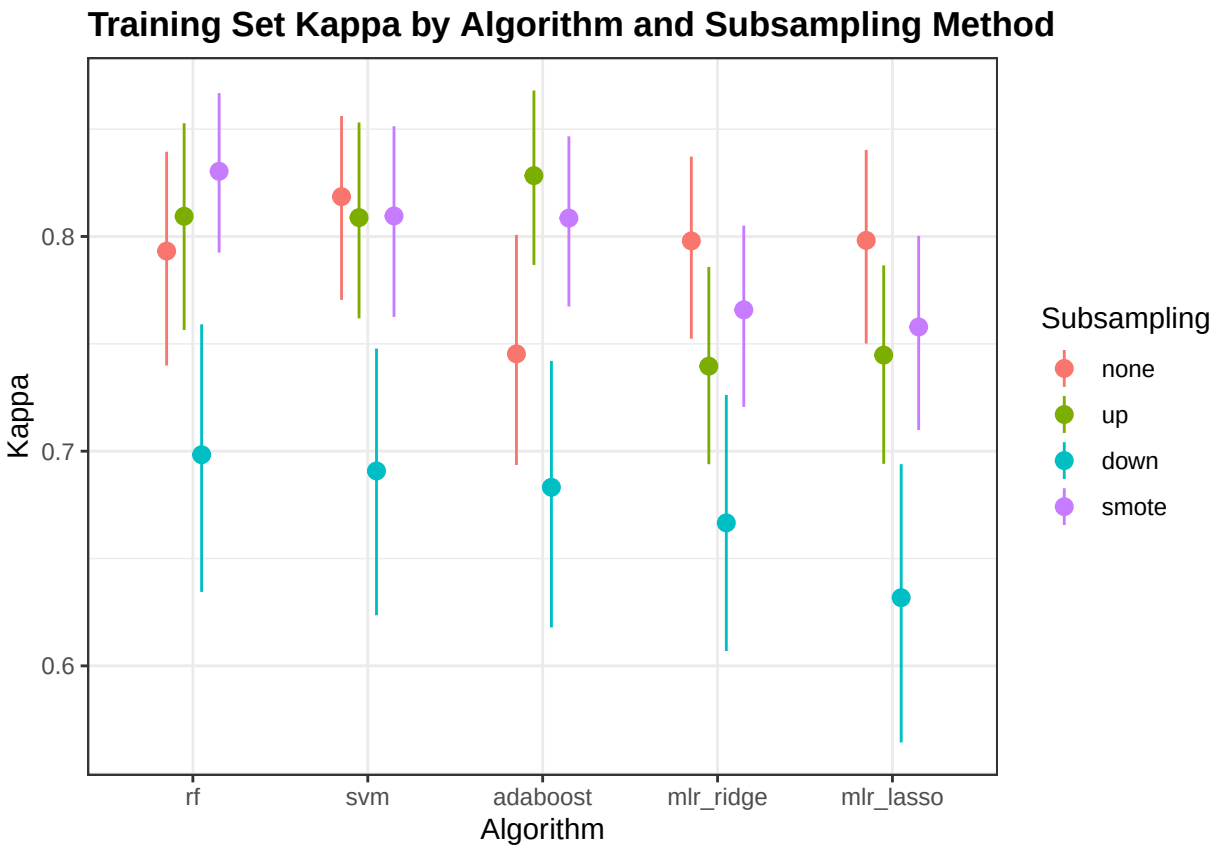

Figure 4.5: Training Set Kappa

Table 4.5: Training Set Kappa by Algorithm and Subsampling Method

sampling	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	0.793	0.819	0.745	0.798	0.798
up	0.809	0.809	0.828	0.74	0.745
down	0.698	0.691	0.683	0.667	0.632
smote	**0.83**	0.81	0.809	0.766	0.758

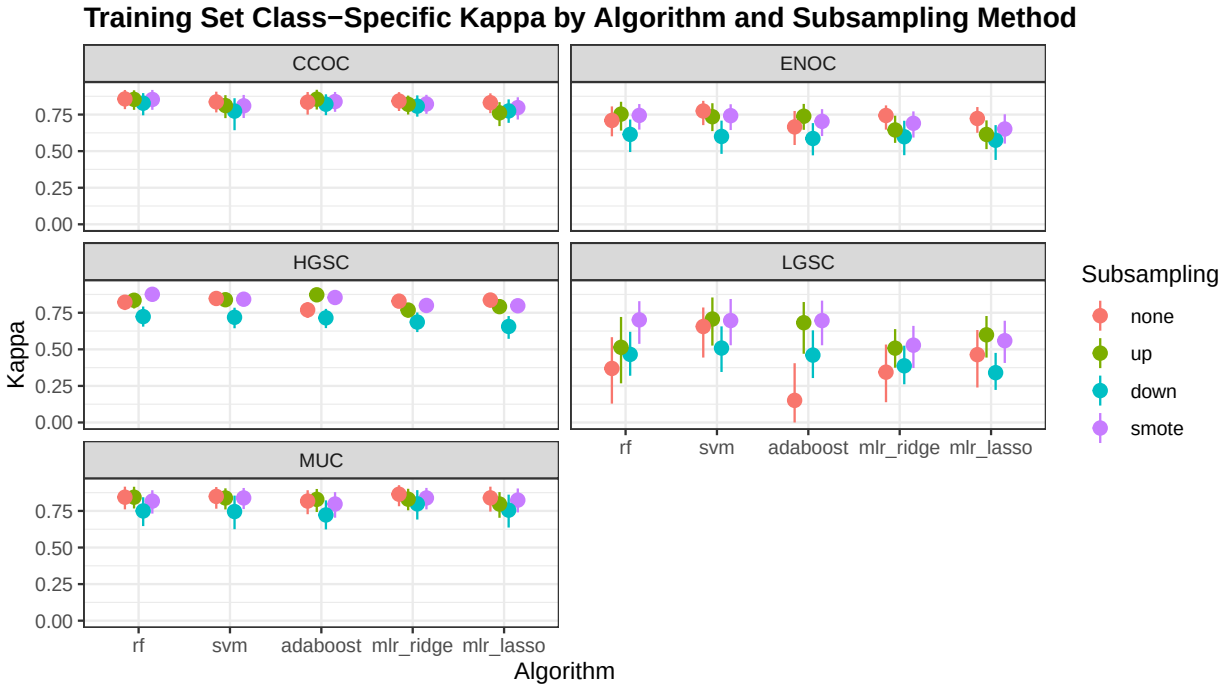

Figure 4.6: Training Set Class-Specific Kappa

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

sampling	histotype	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	CCOC	0.857	0.836	0.835	0.843	0.831
none	ENOC	0.71	0.774	0.666	0.744	0.723
none	HGSC	0.822	0.848	0.77	0.829	0.837
none	LGSC	0.369	0.656	0.151	0.344	0.464
none	MUC	0.844	0.849	0.818	0.864	0.839
up	CCOC	0.854	0.812	0.855	0.82	0.762
up	ENOC	0.753	0.735	0.739	0.645	0.614
up	HGSC	0.835	0.839	0.873	0.768	0.792
up	LGSC	0.514	0.707	0.682	0.508	0.599
up	MUC	0.844	0.839	0.829	0.83	0.797
down	CCOC	0.828	0.774	0.821	0.807	0.775
down	ENOC	0.615	0.6	0.586	0.598	0.575
down	HGSC	0.724	0.719	0.715	0.686	0.656
down	LGSC	0.465	0.509	0.461	0.387	0.341
down	MUC	0.75	0.746	0.723	0.798	0.755
smote	CCOC	0.853	0.81	0.839	0.822	0.798
smote	ENOC	0.744	0.743	0.704	0.69	0.651
smote	HGSC	**0.876**	0.842	0.854	0.8	0.798
smote	LGSC	0.701	0.697	0.696	0.528	0.559
smote	MUC	0.817	0.839	0.797	0.839	0.824

4.1.4 G-mean

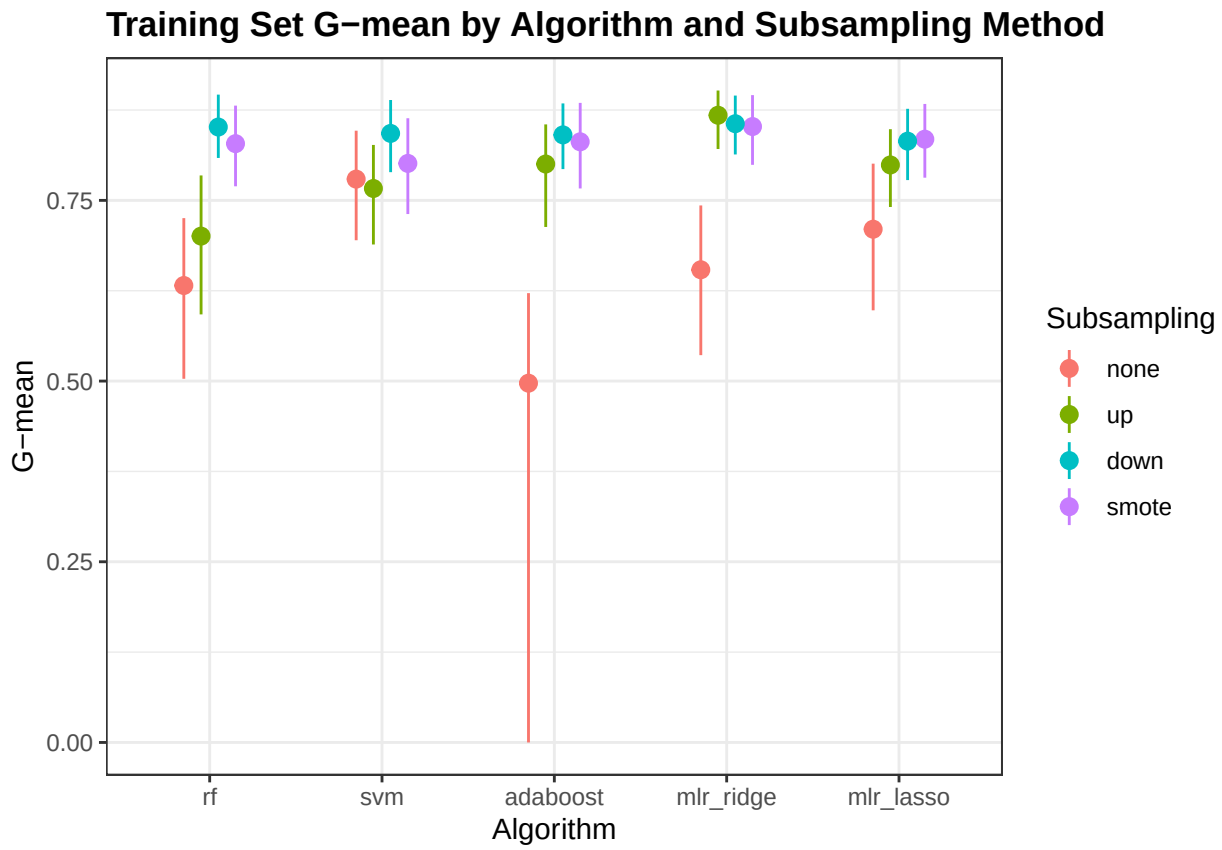

Figure 4.7: Training Set G-mean

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

sampling	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	0.632	0.779	0.497	0.654	0.71
up	0.701	0.766	0.8	**0.868**	0.799
down	0.851	0.843	0.841	0.856	0.832
smote	0.828	0.801	0.831	0.852	0.835

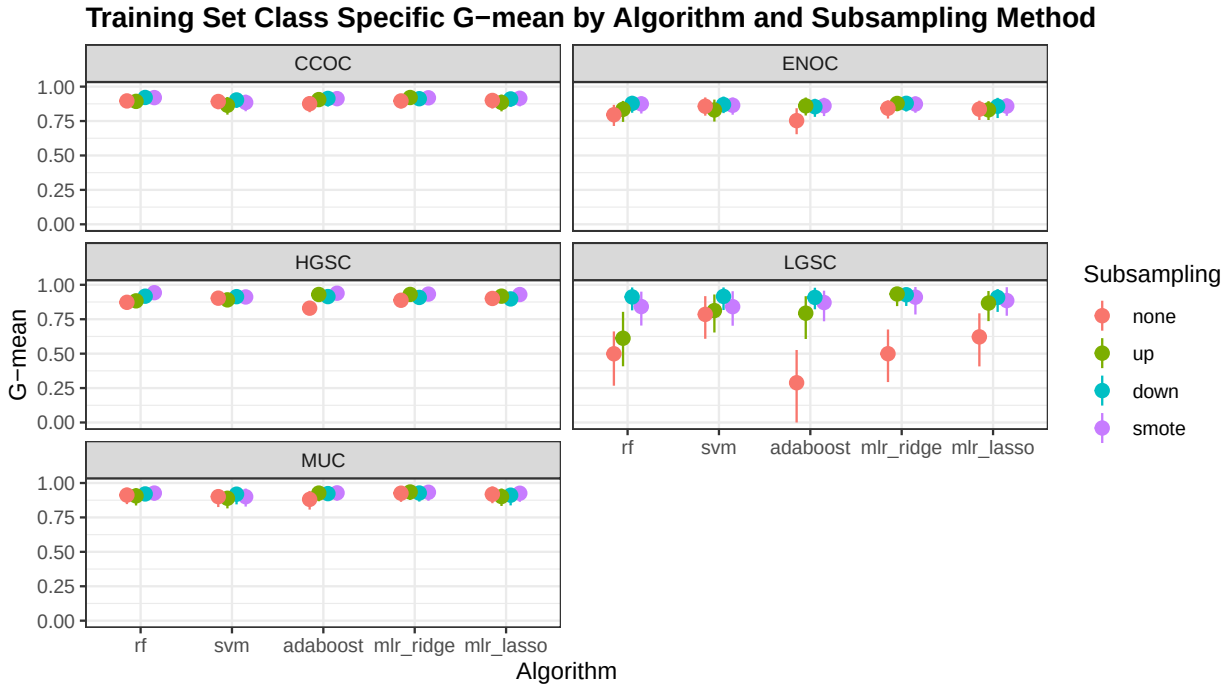

Figure 4.8: Training Set Class-Specific G-mean

4.2 Two-Step Training Set

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

sampling	histotype	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	CCOC	0.896	0.892	0.876	0.896	0.9
none	ENOC	0.796	0.857	0.753	0.843	0.836
none	HGSC	0.873	0.903	0.83	0.888	0.901
none	LGSC	0.499	0.785	0.289	0.5	0.622
none	MUC	0.912	0.901	0.881	0.926	0.918
up	CCOC	0.893	0.864	0.906	0.921	0.885
up	ENOC	0.835	0.83	0.86	0.878	0.831
up	HGSC	0.883	0.891	0.93	0.93	0.918
up	LGSC	0.612	0.813	0.794	0.933	0.867
up	MUC	0.906	0.889	0.927	0.935	0.902
down	CCOC	0.921	0.903	0.914	0.912	0.911
down	ENOC	0.878	0.869	0.854	0.879	0.858
down	HGSC	0.917	0.914	0.914	0.908	0.898
down	LGSC	0.912	0.916	0.909	0.928	0.908
down	MUC	0.921	0.919	0.923	0.927	0.912
smote	CCOC	0.92	0.885	0.913	0.919	0.916
smote	ENOC	0.875	0.865	0.861	0.874	0.858
smote	HGSC	**0.943**	0.912	0.939	0.933	0.929
smote	LGSC	0.842	0.841	0.871	0.91	0.886
smote	MUC	0.927	0.9	0.928	0.933	0.926

4.2.1 Accuracy

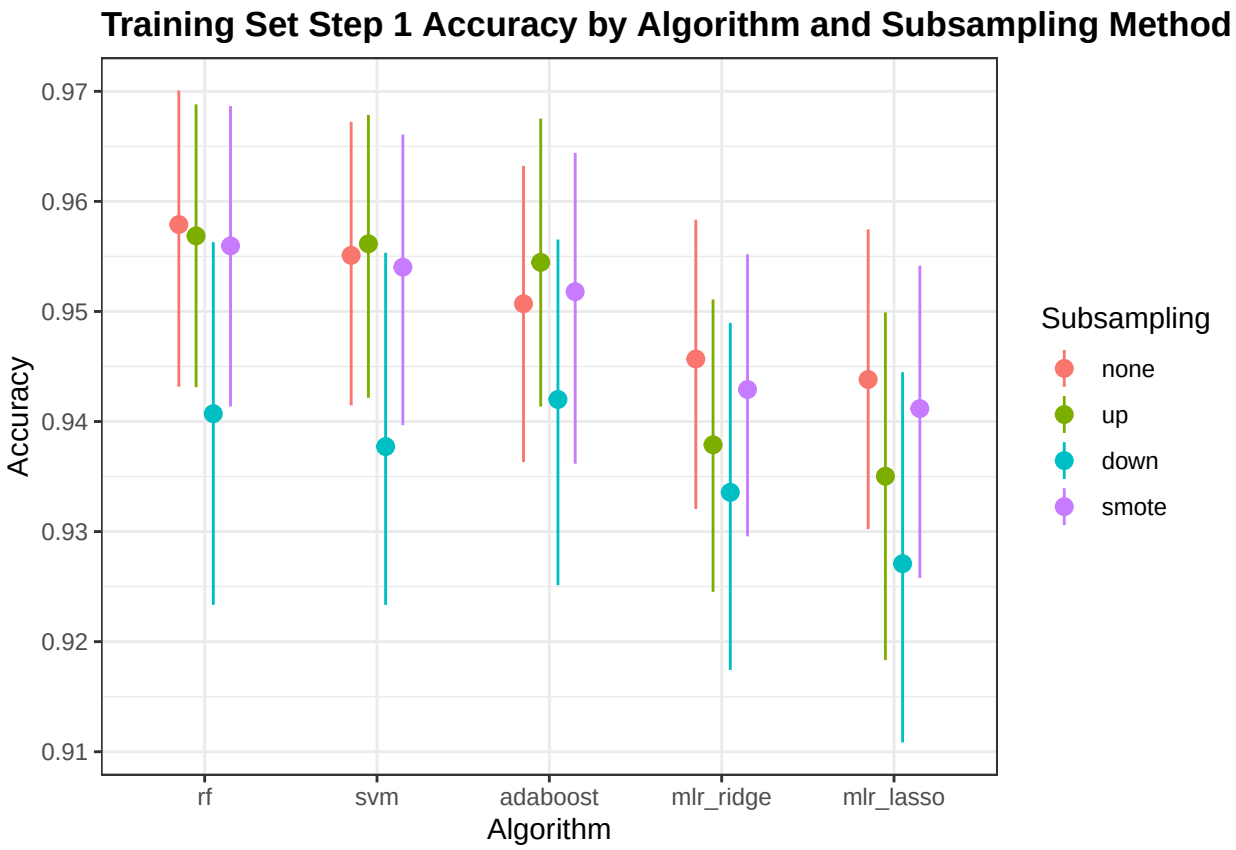

Figure 4.9: Training Set Step 1 Accuracy

Table 4.9: Training Set Step 1 Accuracy by Algorithm and Subsampling Method

sampling	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	**0.958**	0.955	0.951	0.946	0.944
up	0.957	0.956	0.954	0.938	0.935
down	0.941	0.938	0.942	0.934	0.927
smote	0.956	0.954	0.952	0.943	0.941

Table 4.10: Training Set Step 2 Accuracy by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.879	0.876	**0.881**	0.869	0.868
up	**0.881**	0.876	0.877	0.868	0.867
down	0.866	0.862	0.866	0.86	0.854
smote	0.877	0.874	0.873	0.867	0.866

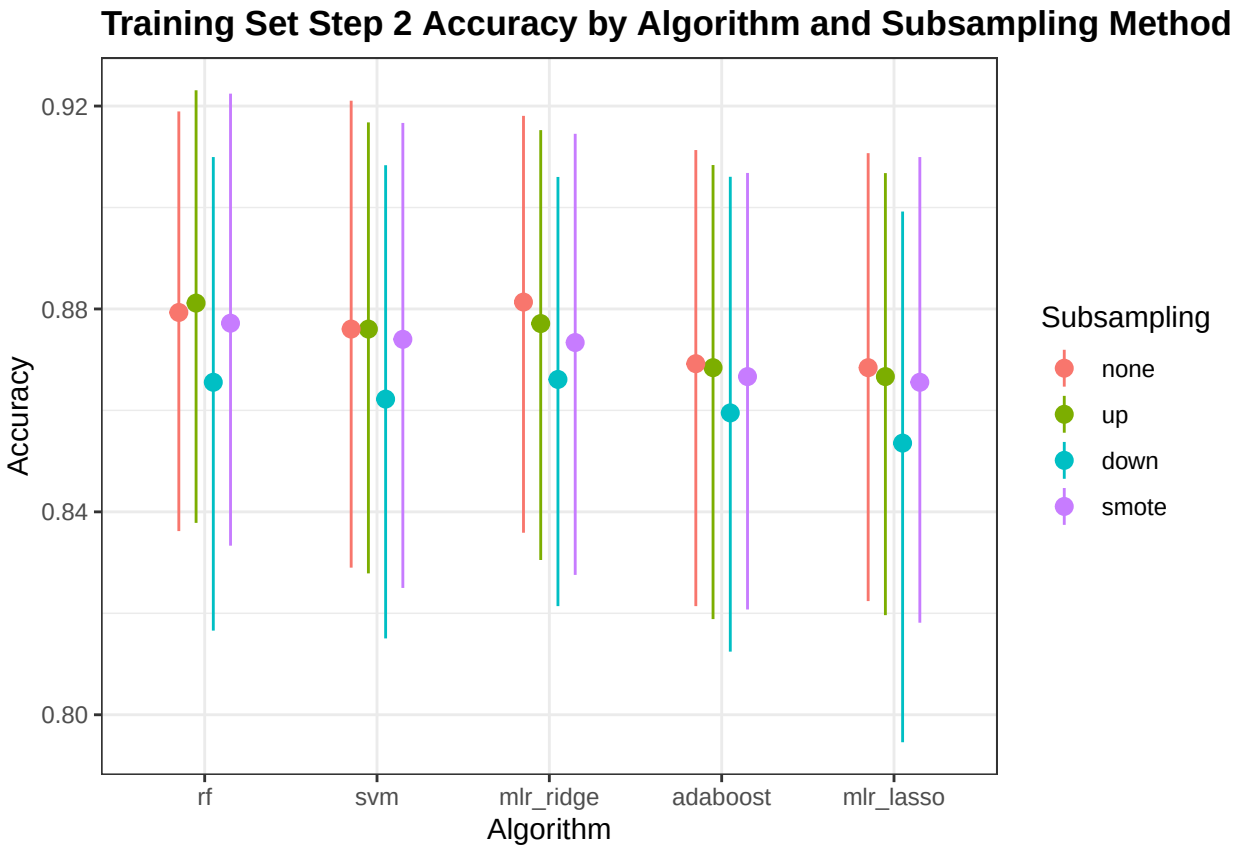

Figure 4.10: Training Set Step 2 Accuracy

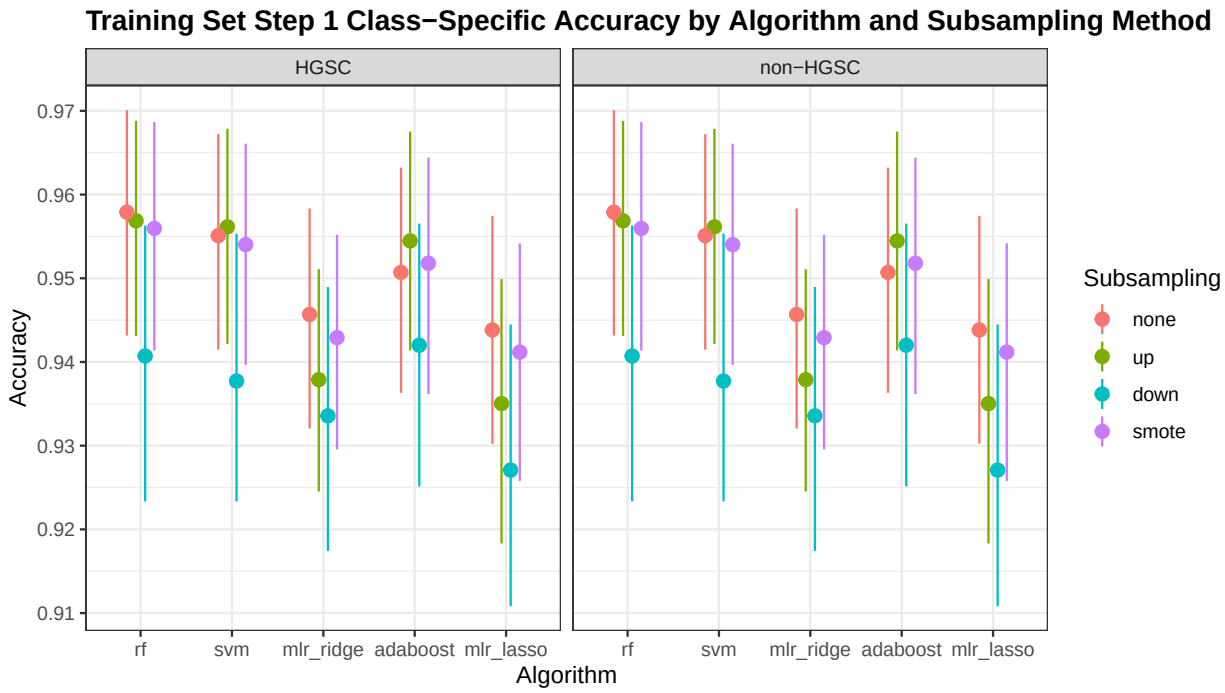

Figure 4.11: Training Set Step 1 Class-Specific Accuracy

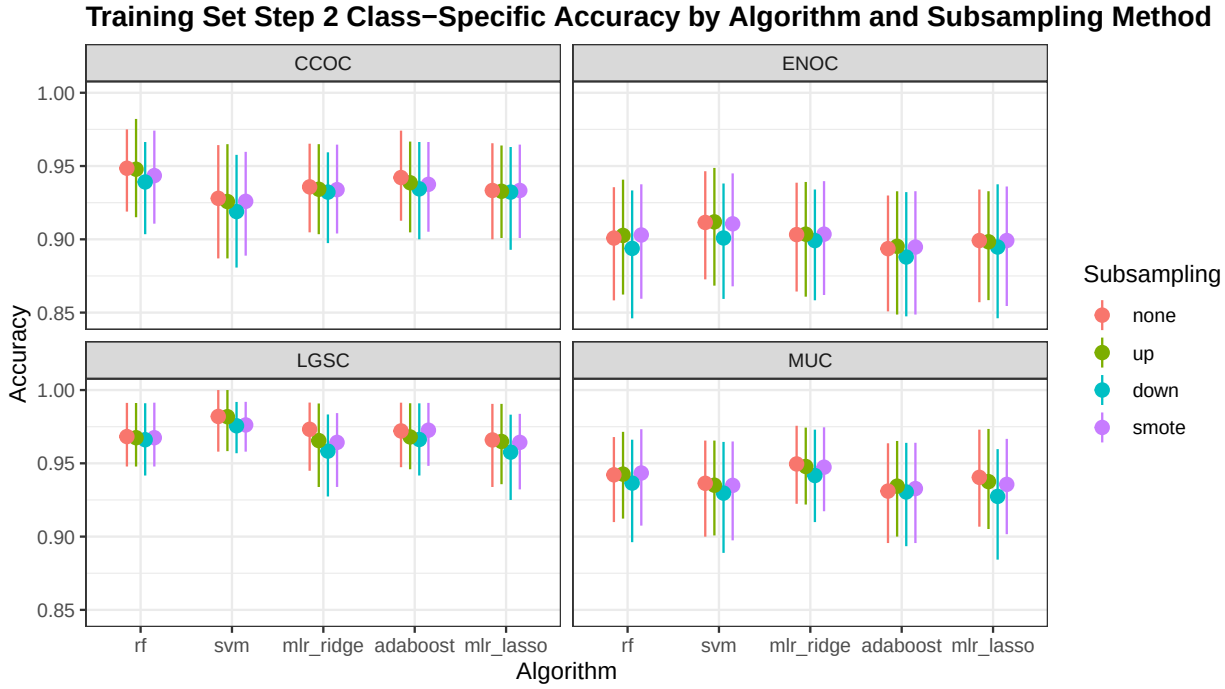

Figure 4.12: Training Set Step 2 Class-Specific Accuracy

Table 4.11: Training Set Step 1 Class-Specific Accuracy by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	HGSC	**0.958**	0.955	0.946	0.951	0.944
none	non-HGSC	**0.958**	0.955	0.946	0.951	0.944
up	HGSC	0.957	0.956	0.938	0.954	0.935
up	non-HGSC	0.957	0.956	0.938	0.954	0.935
down	HGSC	0.941	0.938	0.934	0.942	0.927
down	non-HGSC	0.941	0.938	0.934	0.942	0.927
smote	HGSC	0.956	0.954	0.943	0.952	0.941
smote	non-HGSC	0.956	0.954	0.943	0.952	0.941

Table 4.12: Training Set Step 2 Class-Specific Accuracy by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.948	0.928	0.936	0.942	0.933
none	ENOC	0.901	0.912	0.903	0.894	0.899
none	LGSC	0.968	**0.982**	0.973	0.972	0.966
none	MUC	0.942	0.936	0.95	0.931	0.94
up	CCOC	0.948	0.926	0.934	0.939	0.933
up	ENOC	0.903	0.912	0.904	0.895	0.898
up	LGSC	0.967	**0.982**	0.966	0.968	0.965
up	MUC	0.943	0.935	0.948	0.934	0.938
down	CCOC	0.939	0.919	0.932	0.934	0.932
down	ENOC	0.894	0.901	0.899	0.888	0.895
down	LGSC	0.966	0.976	0.958	0.966	0.958
down	MUC	0.937	0.93	0.942	0.93	0.927
smote	CCOC	0.943	0.926	0.934	0.938	0.933
smote	ENOC	0.903	0.911	0.904	0.895	0.899
smote	LGSC	0.967	0.976	0.964	0.973	0.964
smote	MUC	0.943	0.935	0.947	0.933	0.936

4.2.2 F1-Score

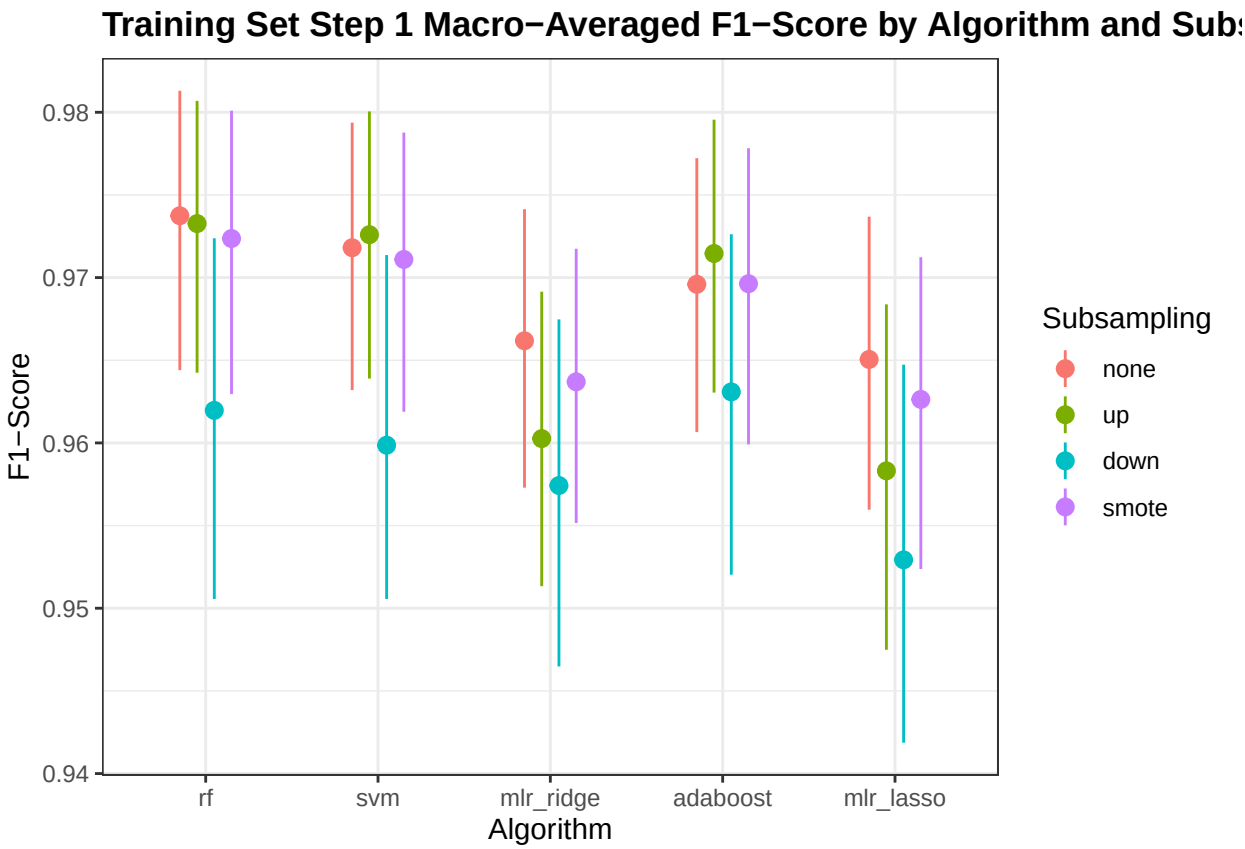

Figure 4.13: Training Set Step 1 F1-Score

Table 4.13: Training Set Step 1 Macro-Averaged F1-Score by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	**0.974**	0.972	0.966	0.97	0.965
up	0.973	0.973	0.96	0.971	0.958
down	0.962	0.96	0.957	0.963	0.953
smote	0.972	0.971	0.964	0.97	0.963

Table 4.14: Training Set Step 2 Macro-Averaged F1-Score by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.879	**0.88**	0.879	0.869	0.865
up	0.879	**0.88**	0.873	0.869	0.862
down	0.865	0.867	0.861	0.858	0.848
smote	0.877	0.878	0.871	0.866	0.862

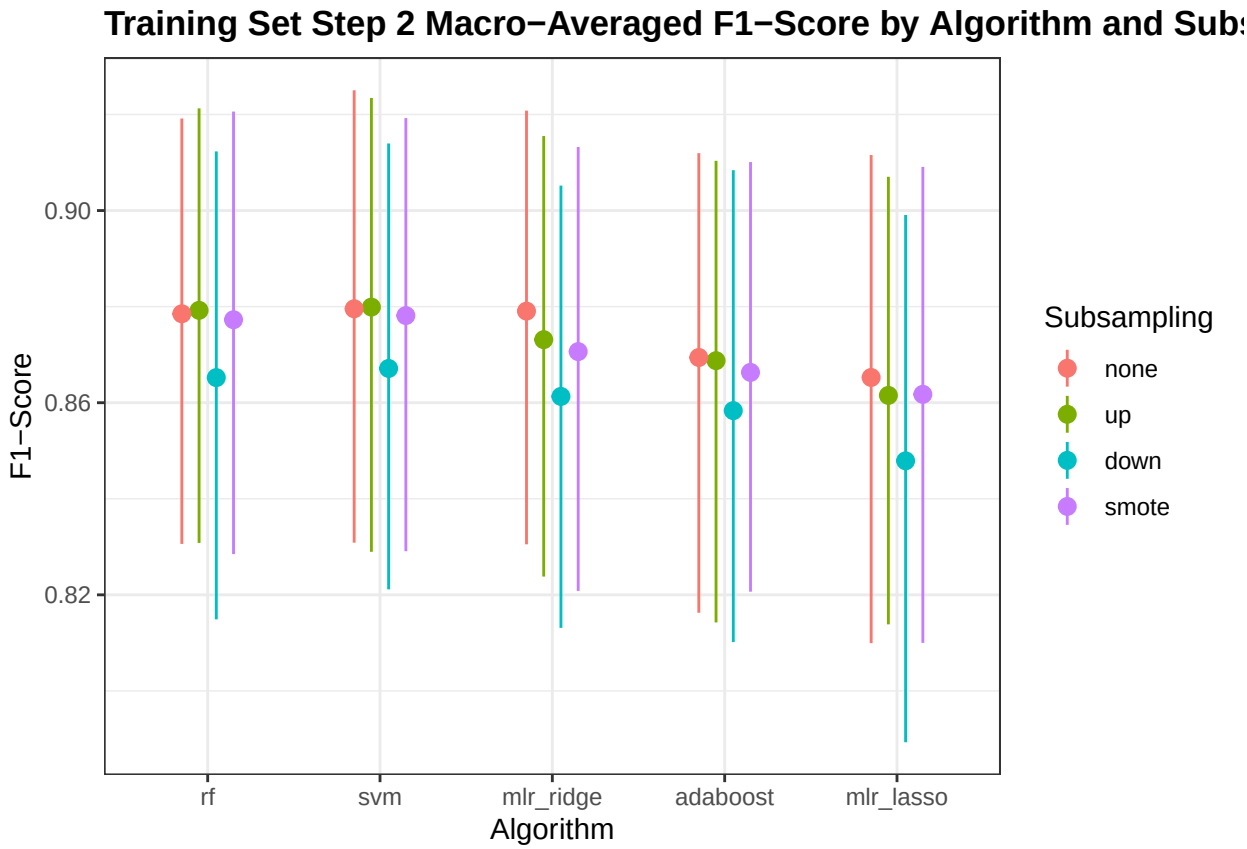

Figure 4.14: Training Set Step 2 F1-Score

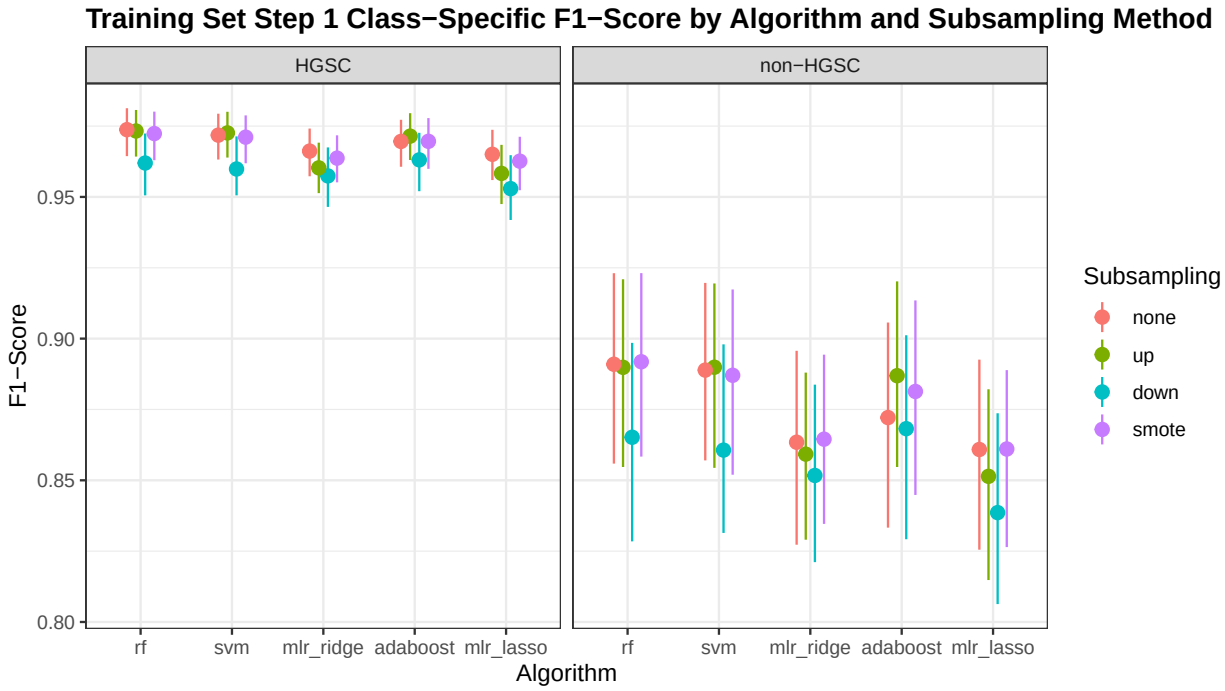

Figure 4.15: Training Set Step 1 Class-Specific F1-Score

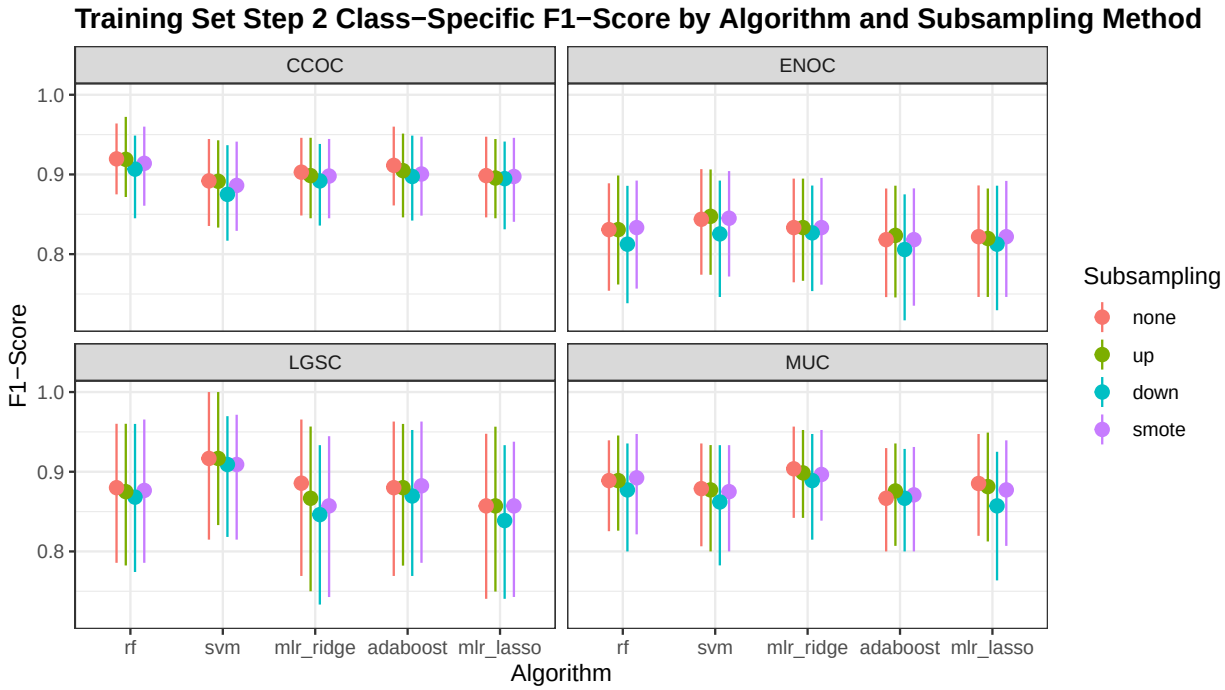

Figure 4.16: Training Set Step 2 Class-Specific F1-Score

Show 50 entries

Search:

Training Set Step 1 Class-Specific F1-Score by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
All	All	All	All	All	All	All
none	HGSC	\textbf{0.974}	0.972	0.966	0.97	0.965
none	non-HGSC	0.891	0.889	0.863	0.872	0.861
up	HGSC	0.973	0.973	0.96	0.971	0.958
up	non-HGSC	0.89	0.89	0.859	0.887	0.851
down	HGSC	0.962	0.96	0.957	0.963	0.953
down	non-HGSC	0.865	0.861	0.852	0.868	0.839
smote	HGSC	0.972	0.971	0.964	0.97	0.963
smote	non-HGSC	0.892	0.887	0.865	0.881	0.861

Showing 1 to 8 of 8 entries

Previous 1 Next

Show 50 entries Search:

Training Set Step 2 Class-Specific F1-Score by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
All	All	All	All	All	All	All
none	CCOC	\textbf{0.92}	0.892	0.903	0.911	0.899
none	ENOC	0.831	0.844	0.833	0.818	0.822
none	LGSC	0.88	0.917	0.886	0.88	0.857
none	MUC	0.889	0.879	0.904	0.867	0.885
up	CCOC	0.919	0.891	0.899	0.905	0.896
up	ENOC	0.831	0.847	0.833	0.824	0.82
up	LGSC	0.875	0.917	0.867	0.88	0.857
up	MUC	0.889	0.877	0.899	0.876	0.881
down	CCOC	0.907	0.875	0.892	0.897	0.895
down	ENOC	0.812	0.825	0.827	0.806	0.812
down	LGSC	0.868	0.909	0.846	0.87	0.839
down	MUC	0.877	0.862	0.889	0.867	0.857
smote	CCOC	0.914	0.886	0.898	0.901	0.897
smote	ENOC	0.833	0.845	0.833	0.818	0.822
smote	LGSC	0.877	0.909	0.857	0.882	0.857
smote	MUC	0.892	0.875	0.897	0.871	0.877

Showing 1 to 16 of 16 entries Previous 1 Next

4.2.3 Kappa

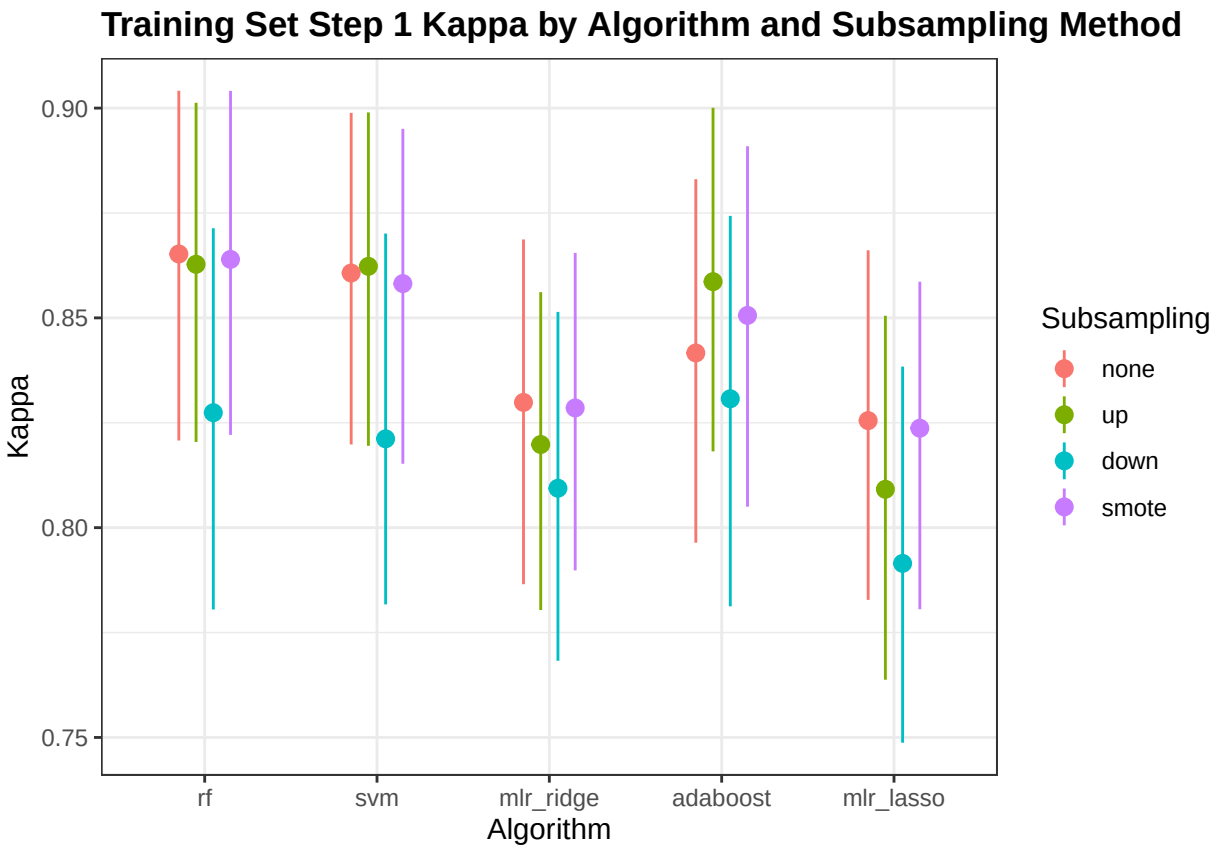

Figure 4.17: Training Set Step 1 Kappa

Table 4.15: Training Set Step 1 Kappa by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	**0.865**	0.861	0.83	0.842	0.826
up	0.863	0.862	0.82	0.859	0.809
down	0.827	0.821	0.809	0.831	0.791
smote	0.864	0.858	0.829	0.851	0.824

Table 4.16: Training Set Step 2 Kappa by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.834	0.829	0.836	0.82	0.819
up	**0.837**	0.828	0.83	0.819	0.816
down	0.815	0.811	0.816	0.806	0.799
smote	0.831	0.826	0.826	0.816	0.815

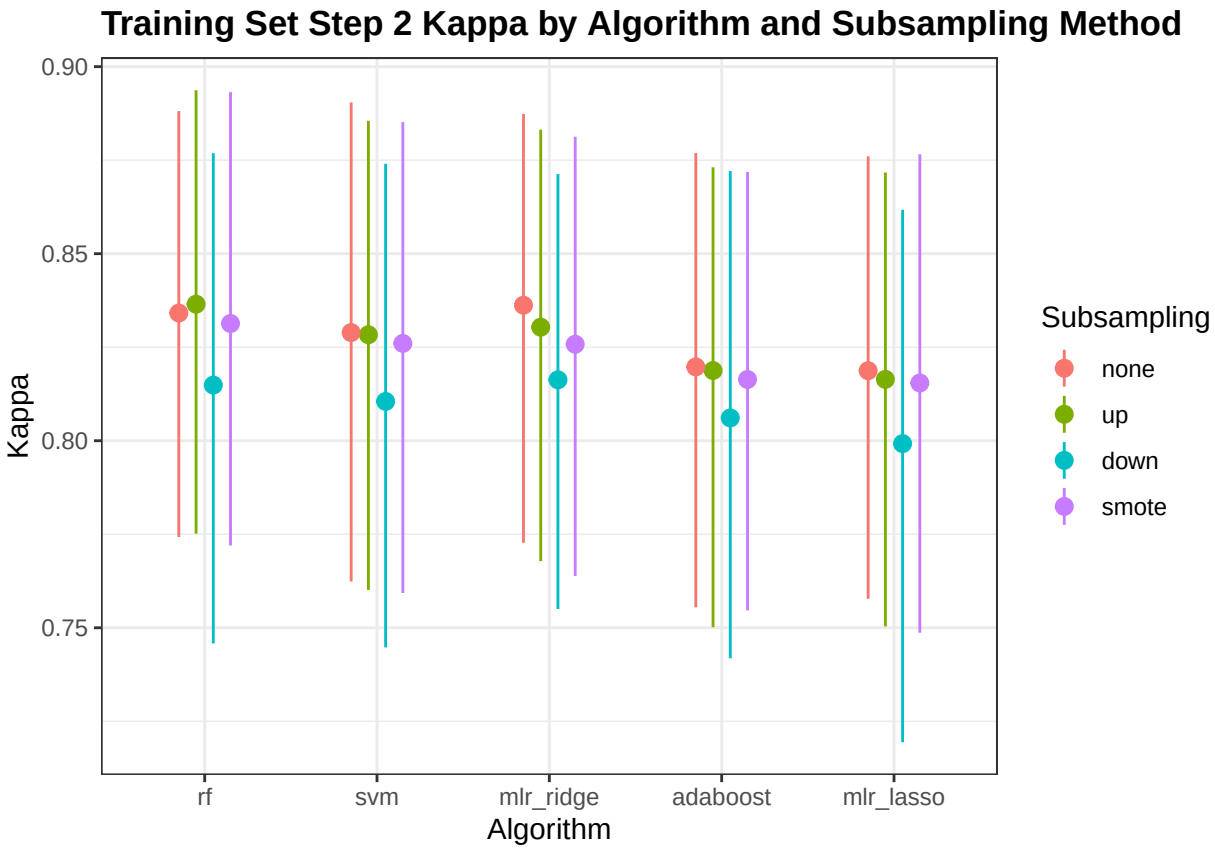

Figure 4.18: Training Set Step 2 Kappa

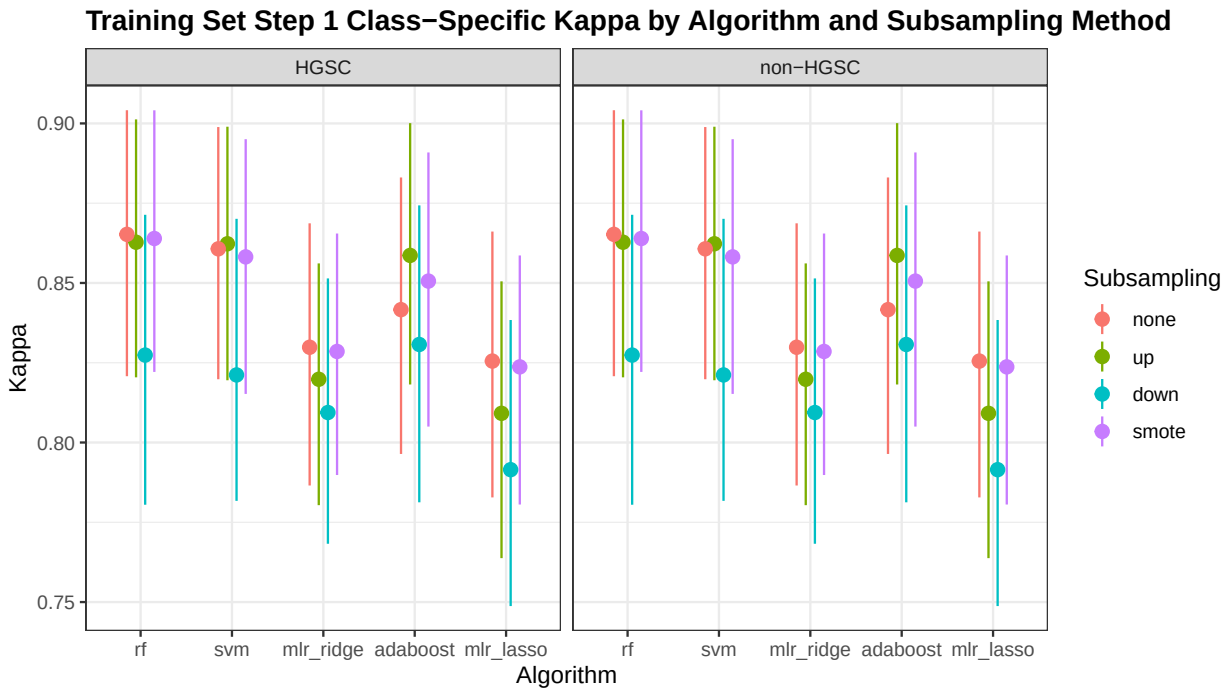

Figure 4.19: Training Set Step 1 Class-Specific Kappa

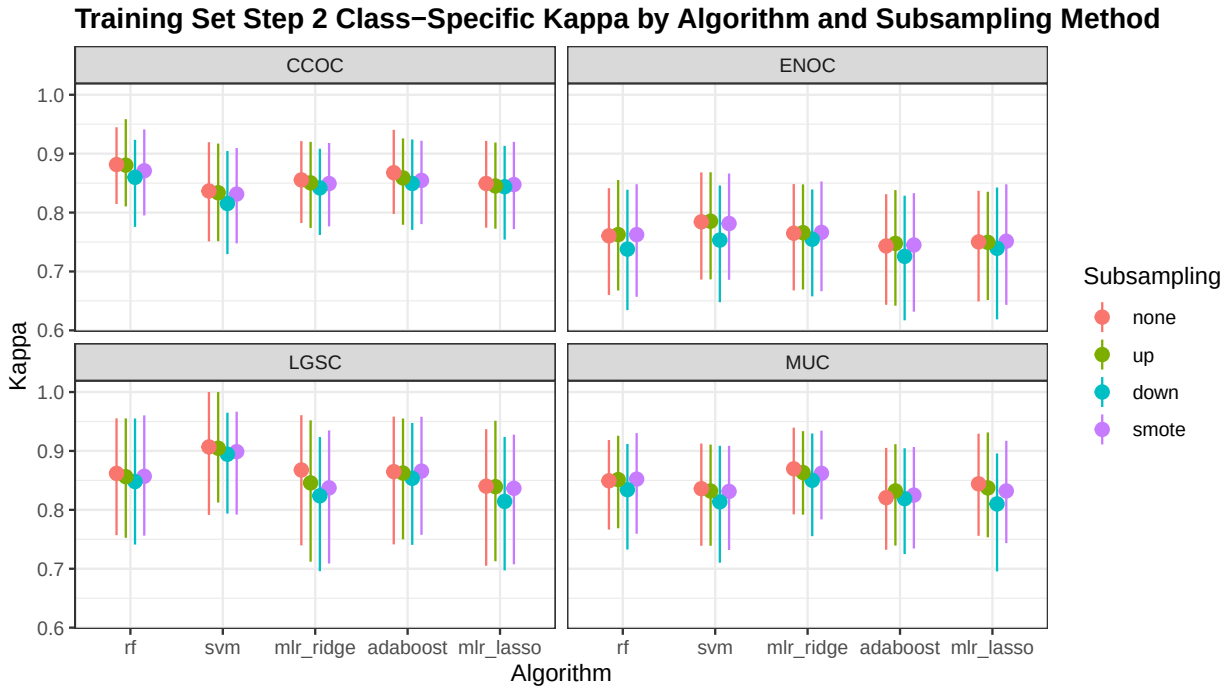

Figure 4.20: Training Set Step 2 Class-Specific Kappa

Table 4.17: Training Set Step 1 Class-Specific Kappa by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	HGSC	**0.865**	0.861	0.83	0.842	0.826
none	non-HGSC	**0.865**	0.861	0.83	0.842	0.826
up	HGSC	0.863	0.862	0.82	0.859	0.809
up	non-HGSC	0.863	0.862	0.82	0.859	0.809
down	HGSC	0.827	0.821	0.809	0.831	0.791
down	non-HGSC	0.827	0.821	0.809	0.831	0.791
smote	HGSC	0.864	0.858	0.829	0.851	0.824
smote	non-HGSC	0.864	0.858	0.829	0.851	0.824

Table 4.18: Training Set Step 2 Class-Specific Kappa by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.882	0.837	0.855	0.868	0.849
none	ENOC	0.76	0.784	0.765	0.743	0.75
none	LGSC	0.862	**0.907**	0.868	0.865	0.84
none	MUC	0.849	0.836	0.87	0.821	0.844
up	CCOC	0.88	0.834	0.85	0.859	0.845
up	ENOC	0.763	0.785	0.766	0.748	0.749
up	LGSC	0.856	0.905	0.846	0.862	0.839
up	MUC	0.851	0.832	0.863	0.832	0.837
down	CCOC	0.86	0.815	0.842	0.849	0.844
down	ENOC	0.738	0.753	0.755	0.726	0.739
down	LGSC	0.848	0.894	0.824	0.854	0.814
down	MUC	0.834	0.814	0.85	0.819	0.81
smote	CCOC	0.871	0.831	0.849	0.855	0.848
smote	ENOC	0.763	0.781	0.766	0.745	0.751
smote	LGSC	0.857	0.899	0.837	0.866	0.837
smote	MUC	0.852	0.831	0.862	0.825	0.832

4.2.4 G-mean

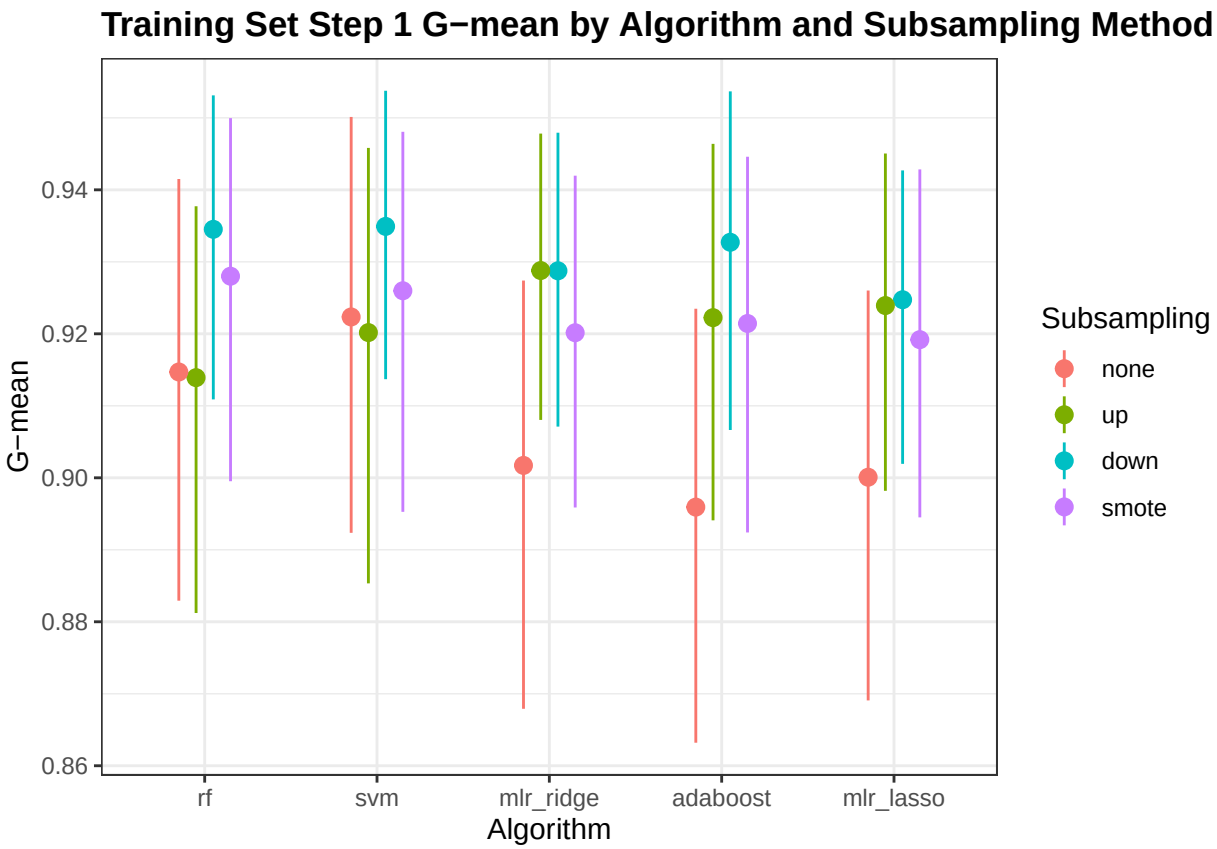

Figure 4.21: Training Set Step 1 G-mean

Table 4.19: Training Set Step 1 G-mean by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.915	0.922	0.902	0.896	0.9
up	0.914	0.92	0.929	0.922	0.924
down	**0.935**	**0.935**	0.929	0.933	0.925
smote	0.928	0.926	0.92	0.921	0.919

Table 4.20: Training Set Step 2 G-mean by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.881	0.878	0.88	0.871	0.867
up	**0.884**	0.878	0.88	0.873	0.868
down	0.873	0.869	0.873	0.865	0.859
smote	0.882	0.875	0.878	0.872	0.87

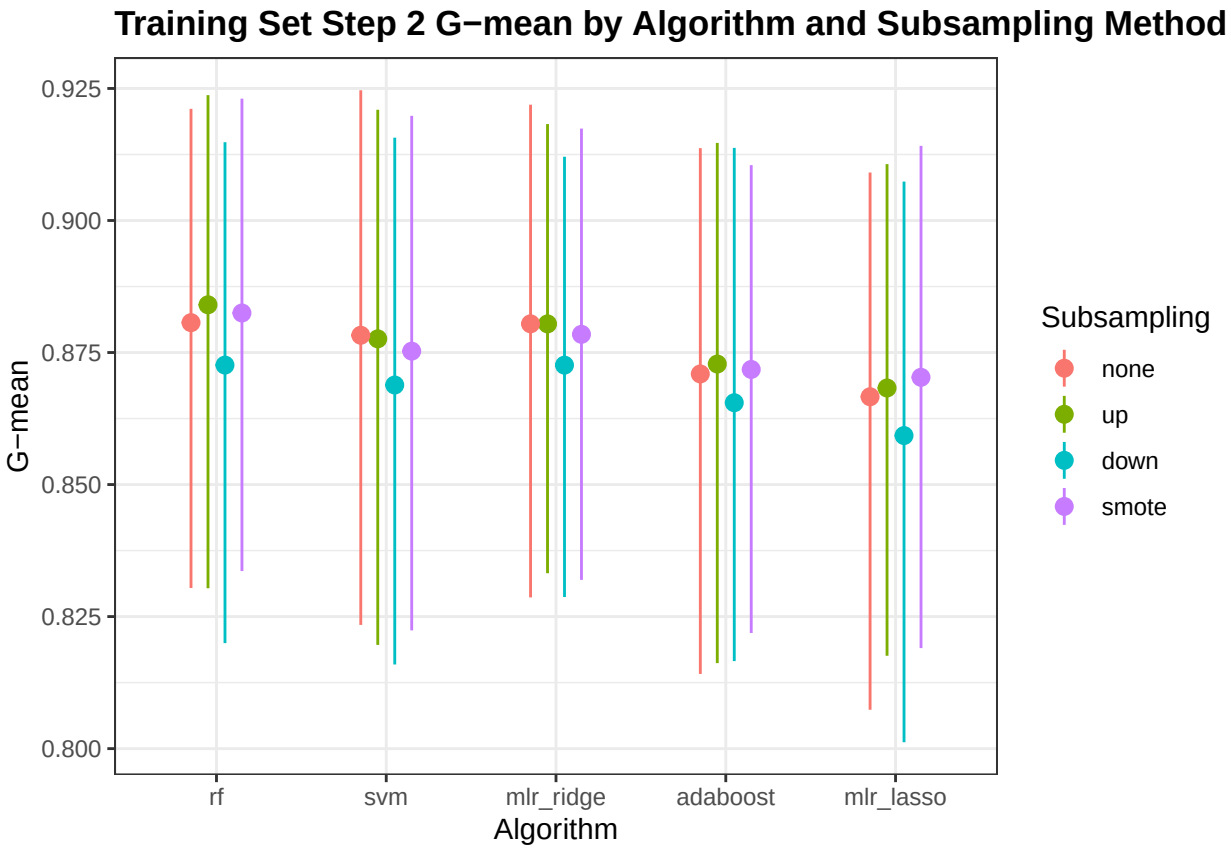

Figure 4.22: Training Set Step 2 G-mean

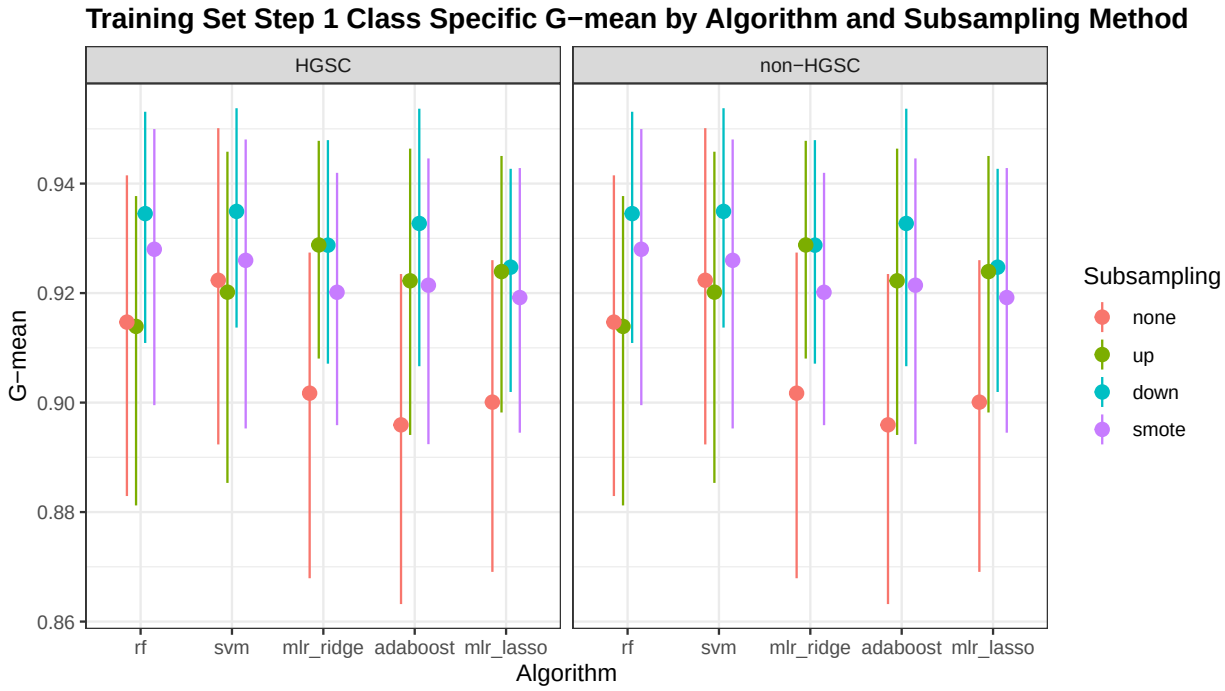

Figure 4.23: Training Set Step 1 Class-Specific G-mean

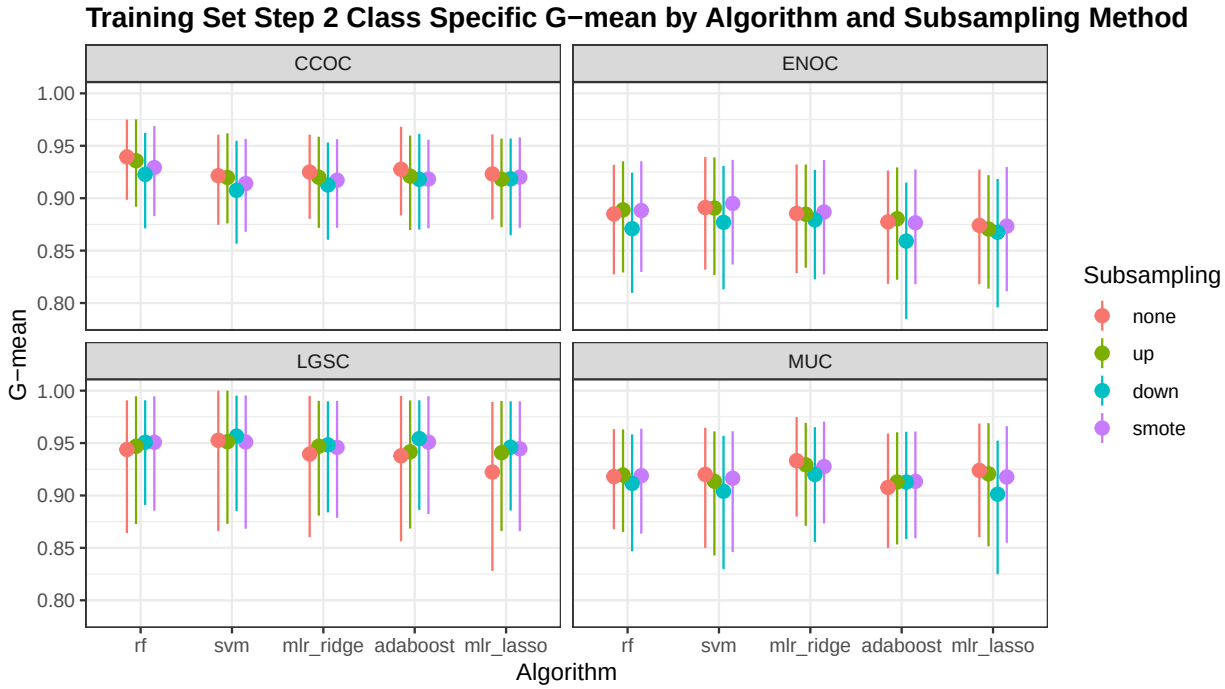

Figure 4.24: Training Set Step 2 Class-Specific G-mean

4.3 CS1 Set

Table 4.21: Training Set Step 1 Class-Specific G-mean by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	HGSC	0.915	0.922	0.902	0.896	0.9
none	non-HGSC	0.915	0.922	0.902	0.896	0.9
up	HGSC	0.914	0.92	0.929	0.922	0.924
up	non-HGSC	0.914	0.92	0.929	0.922	0.924
down	HGSC	**0.935**	**0.935**	0.929	0.933	0.925
down	non-HGSC	**0.935**	**0.935**	0.929	0.933	0.925
smote	HGSC	0.928	0.926	0.92	0.921	0.919
smote	non-HGSC	0.928	0.926	0.92	0.921	0.919

Table 4.22: Training Set Step 2 Class-Specific G-mean by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.939	0.921	0.925	0.928	0.923
none	ENOC	0.885	0.891	0.885	0.877	0.874
none	LGSC	0.944	0.953	0.94	0.938	0.922
none	MUC	0.918	0.92	0.933	0.908	0.924
up	CCOC	0.936	0.92	0.92	0.921	0.918
up	ENOC	0.889	0.891	0.885	0.88	0.871
up	LGSC	0.947	0.951	0.947	0.942	0.941
up	MUC	0.92	0.913	0.929	0.913	0.921
down	CCOC	0.923	0.907	0.913	0.918	0.918
down	ENOC	0.871	0.877	0.879	0.859	0.867
down	LGSC	0.951	**0.957**	0.948	0.954	0.946
down	MUC	0.912	0.904	0.92	0.913	0.901
smote	CCOC	0.929	0.914	0.917	0.918	0.92
smote	ENOC	0.888	0.895	0.887	0.877	0.873
smote	LGSC	0.951	0.951	0.946	0.951	0.945
smote	MUC	0.919	0.916	0.928	0.913	0.918

Table 4.23: CS1 Set Accuracy by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.816	**0.842**	0.832	0.8	0.823
up	0.835	0.833	0.83	0.825	0.814
down	0.798	0.802	0.788	0.776	0.764
smote	0.837	0.837	0.83	0.828	0.814

4.3.1 Accuracy

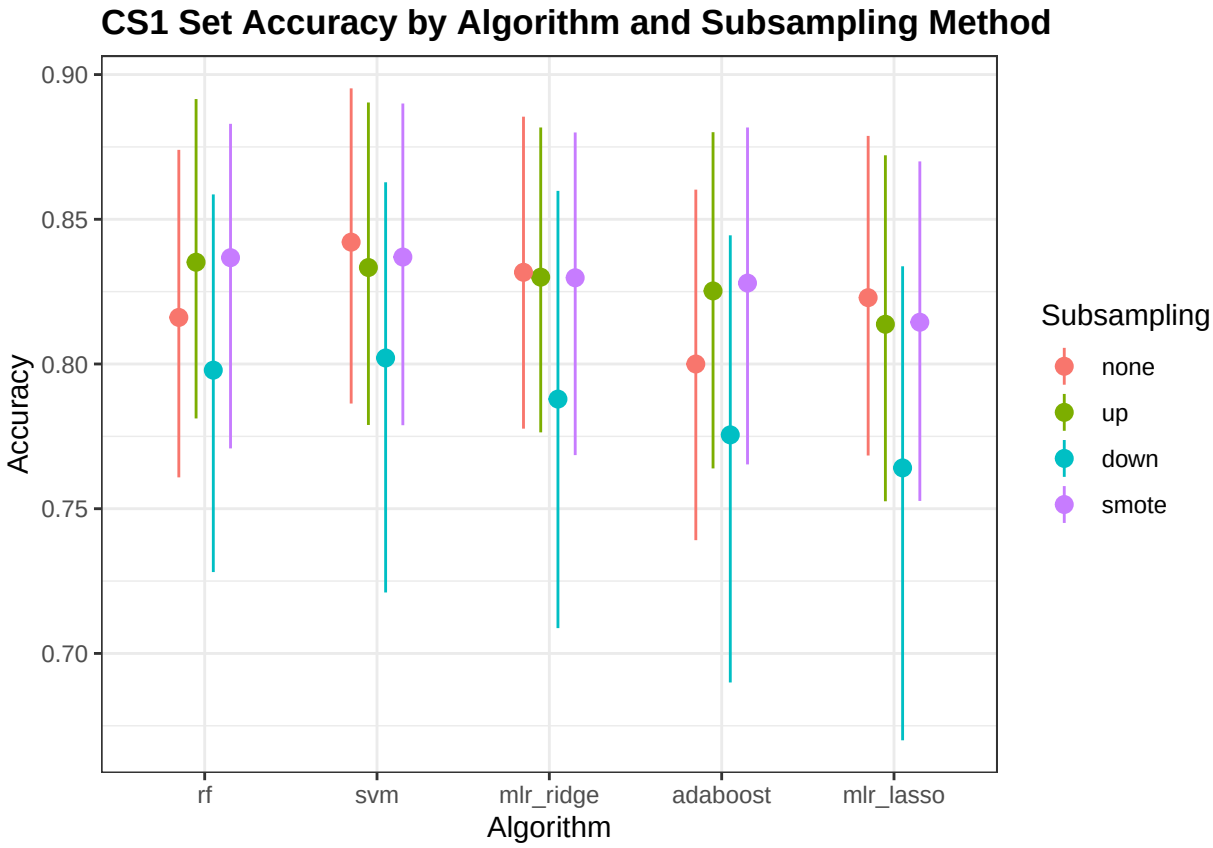

Figure 4.25: CS1 Set Accuracy

Table 4.24: CS1 Set Class-Specific Accuracy by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.939	0.943	0.935	0.939	0.931
none	ENOC	0.894	0.917	0.902	0.892	0.9
none	HGSC	0.891	0.894	0.903	0.87	0.896
none	LGSC	0.95	0.969	0.96	0.943	0.957
none	MUC	0.966	0.962	0.969	0.96	0.969
up	CCOC	0.942	0.938	0.927	0.939	0.918
up	ENOC	0.902	0.906	0.899	0.897	0.885
up	HGSC	0.907	0.894	0.908	0.897	0.903
up	LGSC	0.96	**0.971**	0.959	0.958	0.958
up	MUC	0.967	0.963	0.968	0.961	0.969
down	CCOC	0.936	0.938	0.937	0.932	0.922
down	ENOC	0.887	0.895	0.892	0.876	0.878
down	HGSC	0.883	0.874	0.872	0.869	0.858
down	LGSC	0.939	0.949	0.925	0.935	0.921
down	MUC	0.958	0.954	0.96	0.95	0.957
smote	CCOC	0.94	0.939	0.929	0.939	0.926
smote	ENOC	0.892	0.905	0.896	0.888	0.891
smote	HGSC	0.916	0.897	0.907	0.905	0.894
smote	LGSC	0.961	0.97	0.96	0.963	0.957
smote	MUC	0.962	0.967	0.968	0.96	0.969

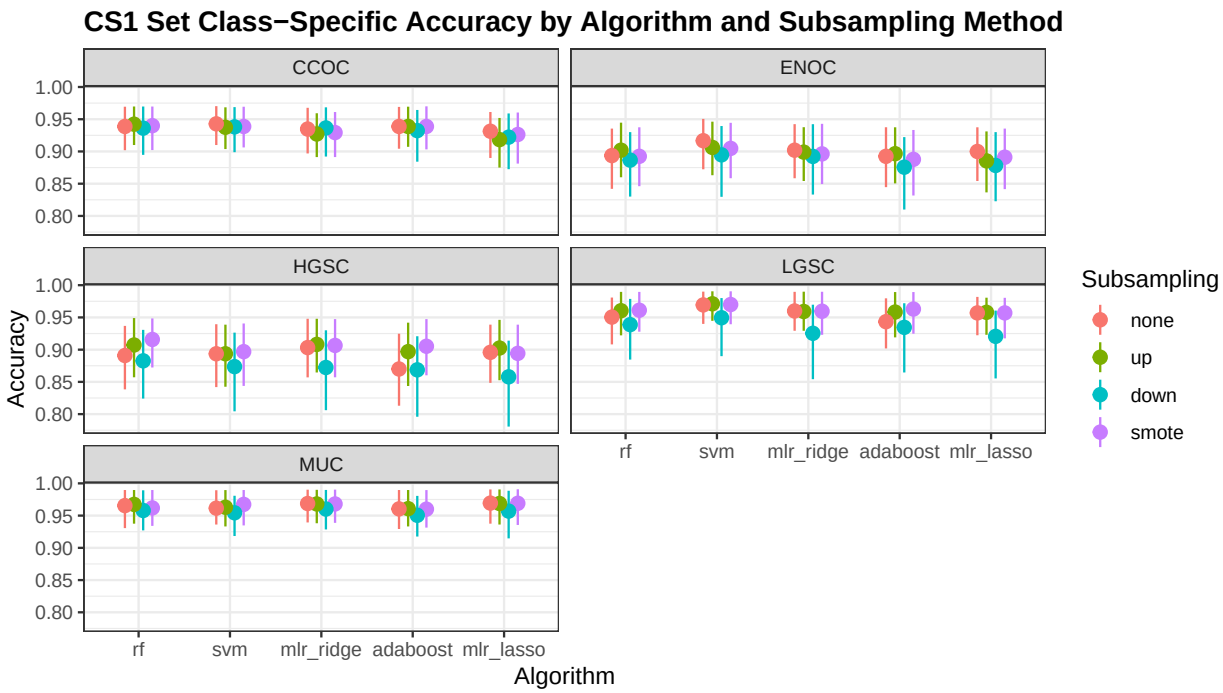

Figure 4.26: CS1 Set Class-Specific Accuracy

Table 4.25: CS1 Set Macro-Averaged F1-Score by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.731	**0.791**	0.771	0.707	0.761
up	0.77	0.781	0.786	0.755	0.77
down	0.746	0.754	0.745	0.728	0.716
smote	0.779	0.789	0.784	0.773	0.773

4.3.2 F1-Score

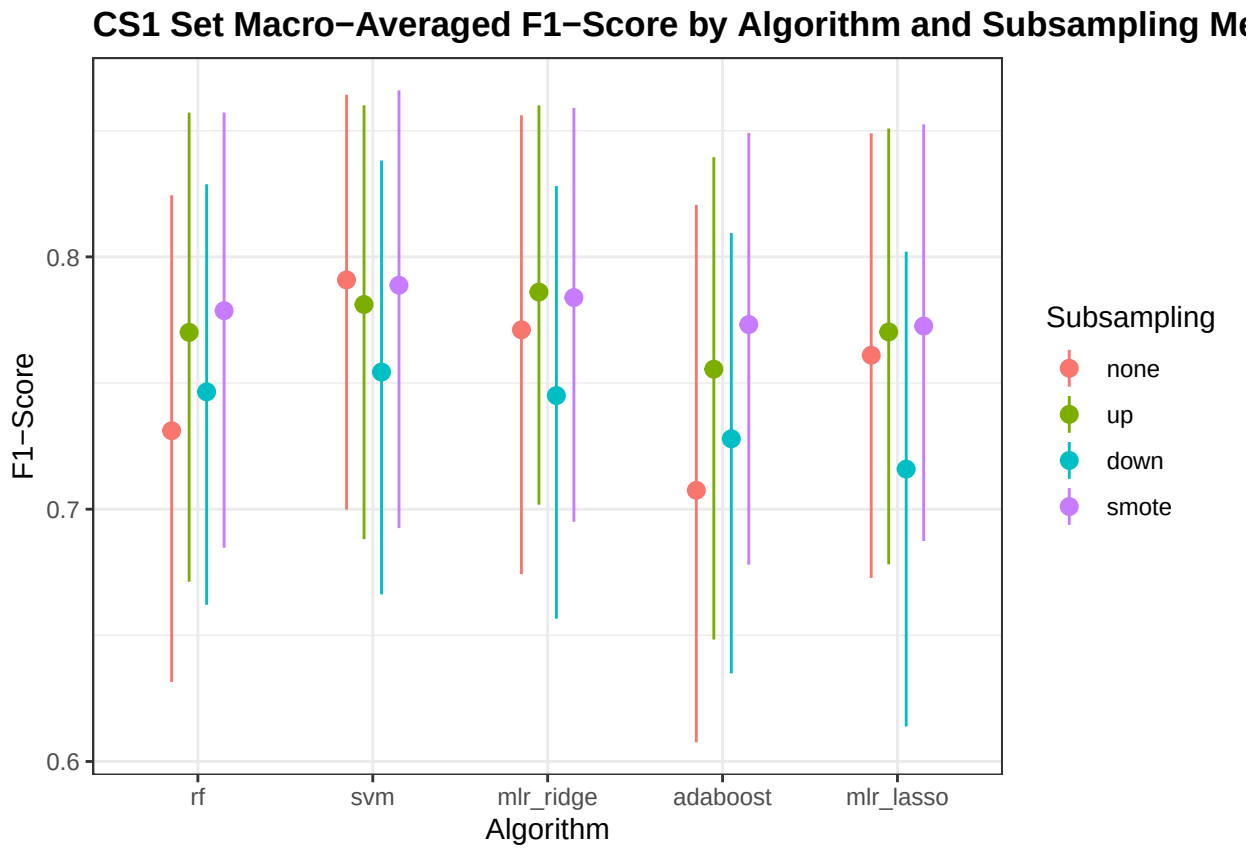

Figure 4.27: CS1 Set F1-Score

Table 4.26: CS1 Set Class-Specific F1-Score by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.824	0.828	0.812	0.824	0.8
none	ENOC	0.762	0.8	0.769	0.745	0.772
none	HGSC	0.889	0.893	0.9	0.873	0.891
none	LGSC	0.5	0.75	0.667	0.4	0.615
none	MUC	0.667	0.714	0.75	0.667	0.75
up	CCOC	0.833	0.815	0.8	0.824	0.773
up	ENOC	0.78	0.783	0.769	0.765	0.744
up	HGSC	0.905	0.893	0.901	0.895	0.891
up	LGSC	0.667	0.769	0.727	0.615	0.714
up	MUC	0.727	0.667	0.75	0.714	0.762
down	CCOC	0.822	0.824	0.821	0.811	0.786
down	ENOC	0.75	0.769	0.762	0.718	0.723
down	HGSC	0.864	0.857	0.851	0.85	0.835
down	LGSC	0.632	0.667	0.588	0.615	0.571
down	MUC	0.706	0.706	0.714	0.667	0.667
smote	CCOC	0.829	0.828	0.81	0.828	0.789
smote	ENOC	0.769	0.783	0.766	0.756	0.757
smote	HGSC	**0.909**	0.892	0.897	0.899	0.884
smote	LGSC	0.714	0.75	0.727	0.714	0.706
smote	MUC	0.727	0.714	0.75	0.706	0.762

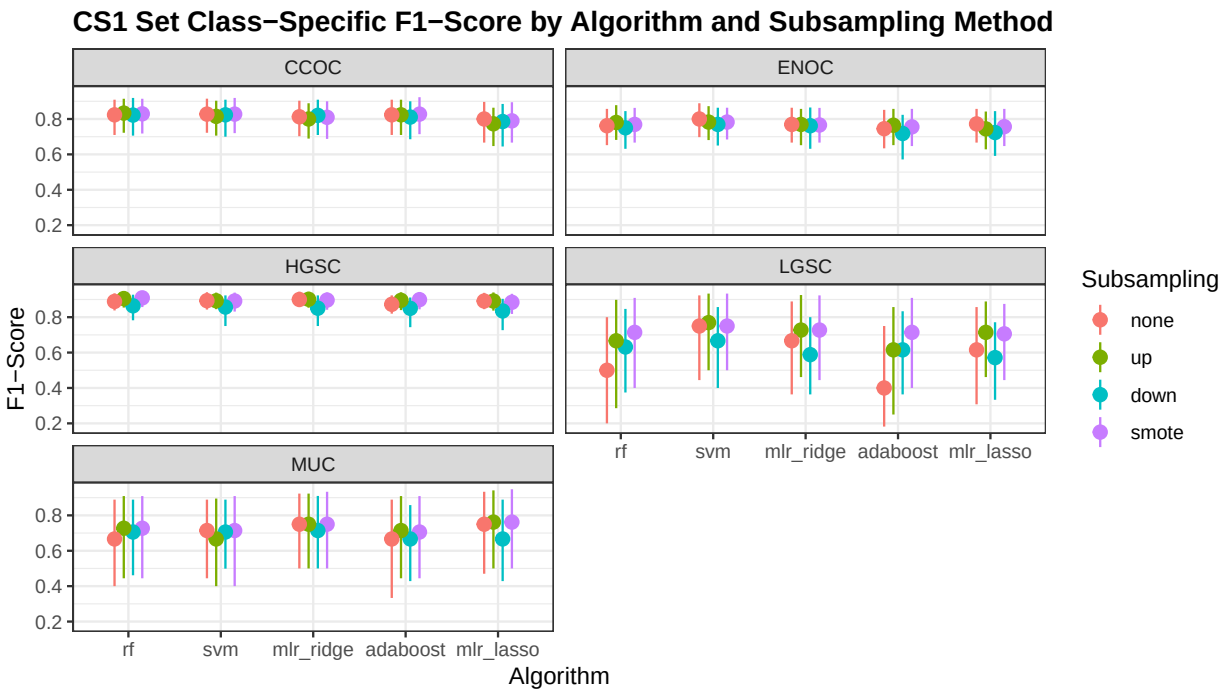

Figure 4.28: CS1 Set Class-Specific F1-Score

Table 4.27: CS1 Set Kappa by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.724	**0.767**	0.752	0.697	0.74
up	0.756	0.752	0.757	0.741	0.732
down	0.716	0.72	0.706	0.687	0.673
smote	0.763	0.759	0.755	0.751	0.736

4.3.3 Kappa

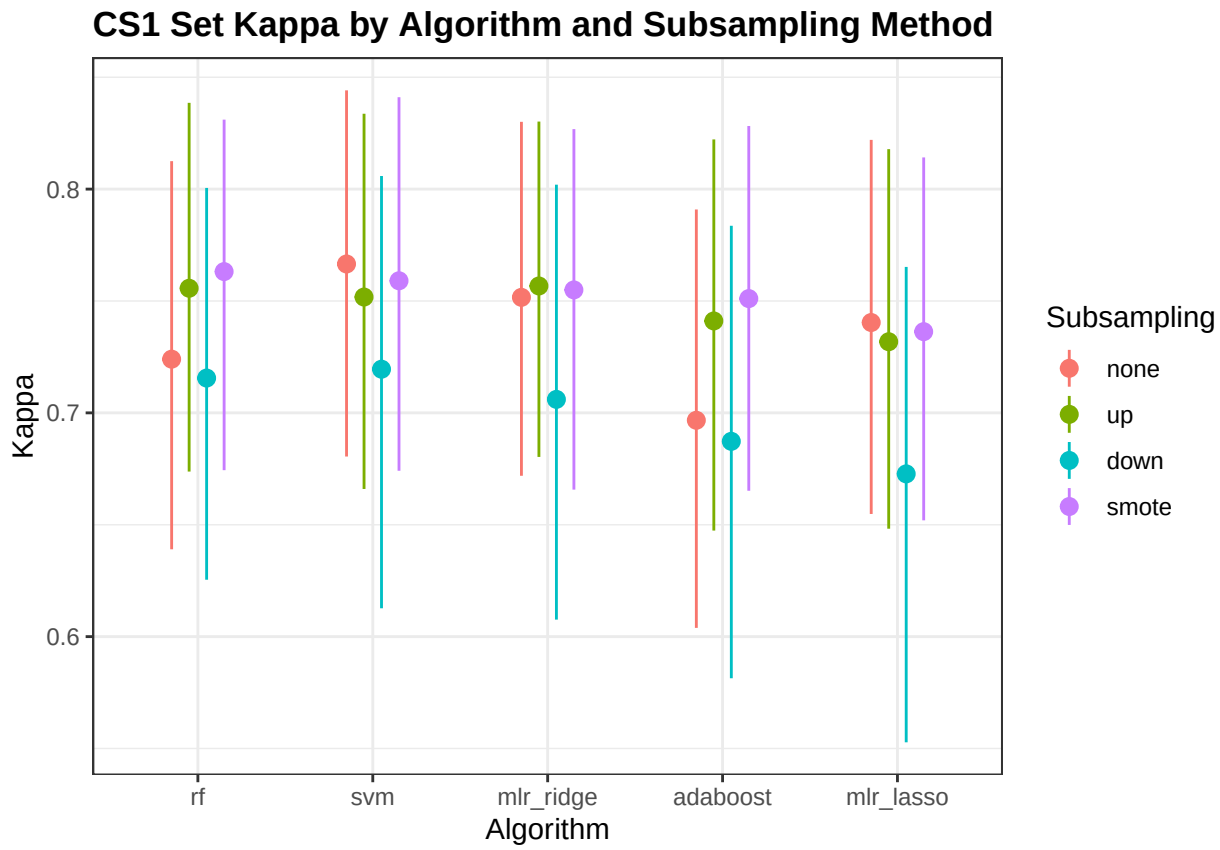

Figure 4.29: CS1 Set Kappa

Table 4.28: CS1 Set Class-Specific Kappa by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.786	0.794	0.773	0.785	0.764
none	ENOC	0.693	0.75	0.707	0.678	0.704
none	HGSC	0.781	0.787	0.806	0.741	0.792
none	LGSC	0.477	0.728	0.646	0.273	0.587
none	MUC	0.652	0.693	0.729	0.647	0.739
up	CCOC	0.797	0.777	0.76	0.787	0.718
up	ENOC	0.718	0.723	0.702	0.699	0.669
up	HGSC	0.814	0.788	0.816	0.795	0.802
up	LGSC	0.647	0.753	0.696	0.585	0.692
up	MUC	0.71	0.647	0.739	0.692	0.74
down	CCOC	0.784	0.786	0.778	0.767	0.735
down	ENOC	0.674	0.701	0.69	0.636	0.647
down	HGSC	0.761	0.742	0.741	0.734	0.711
down	LGSC	0.594	0.64	0.553	0.587	0.528
down	MUC	0.678	0.675	0.693	0.647	0.65
smote	CCOC	0.795	0.79	0.762	0.79	0.747
smote	ENOC	0.699	0.725	0.7	0.678	0.685
smote	HGSC	**0.829**	0.792	0.81	0.81	0.787
smote	LGSC	0.691	0.74	0.711	0.694	0.676
smote	MUC	0.709	0.694	0.73	0.678	0.74

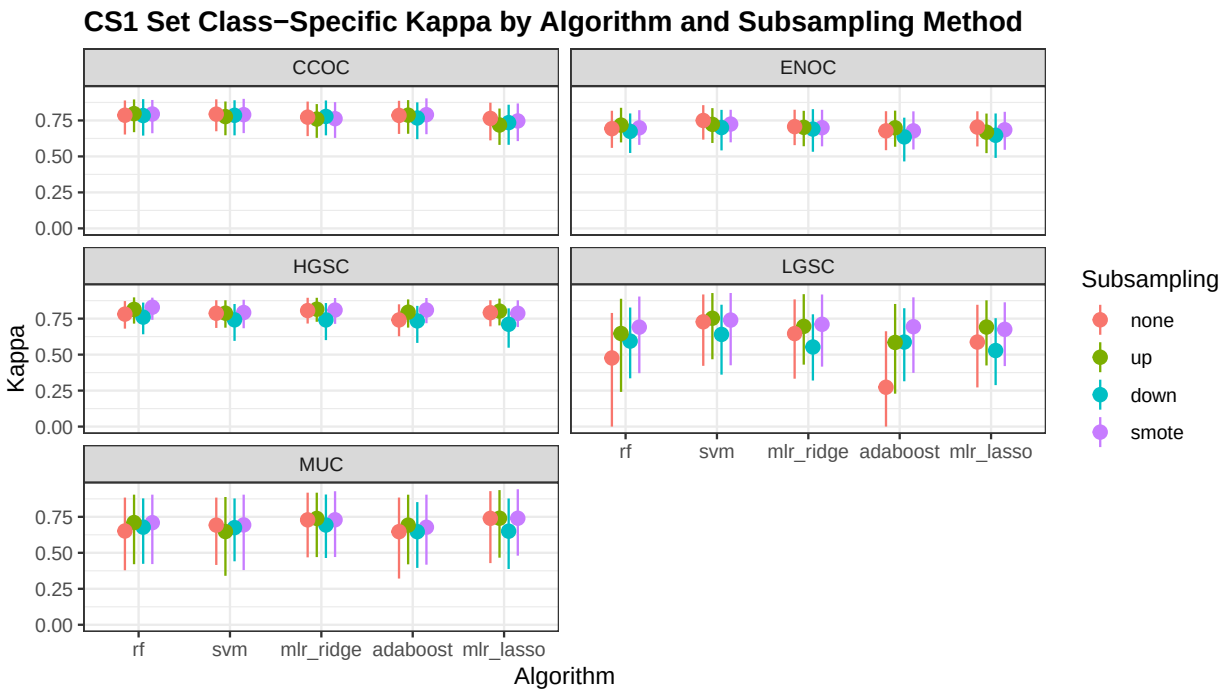

Figure 4.30: CS1 Set Class-Specific Kappa

Table 4.29: CS1 Set G-mean by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.638	0.747	0.728	0.557	0.72
up	0.713	0.718	**0.794**	0.703	0.769
down	0.778	0.791	0.779	0.761	0.753
smote	0.76	0.75	0.789	0.766	0.782

4.3.4 G-mean

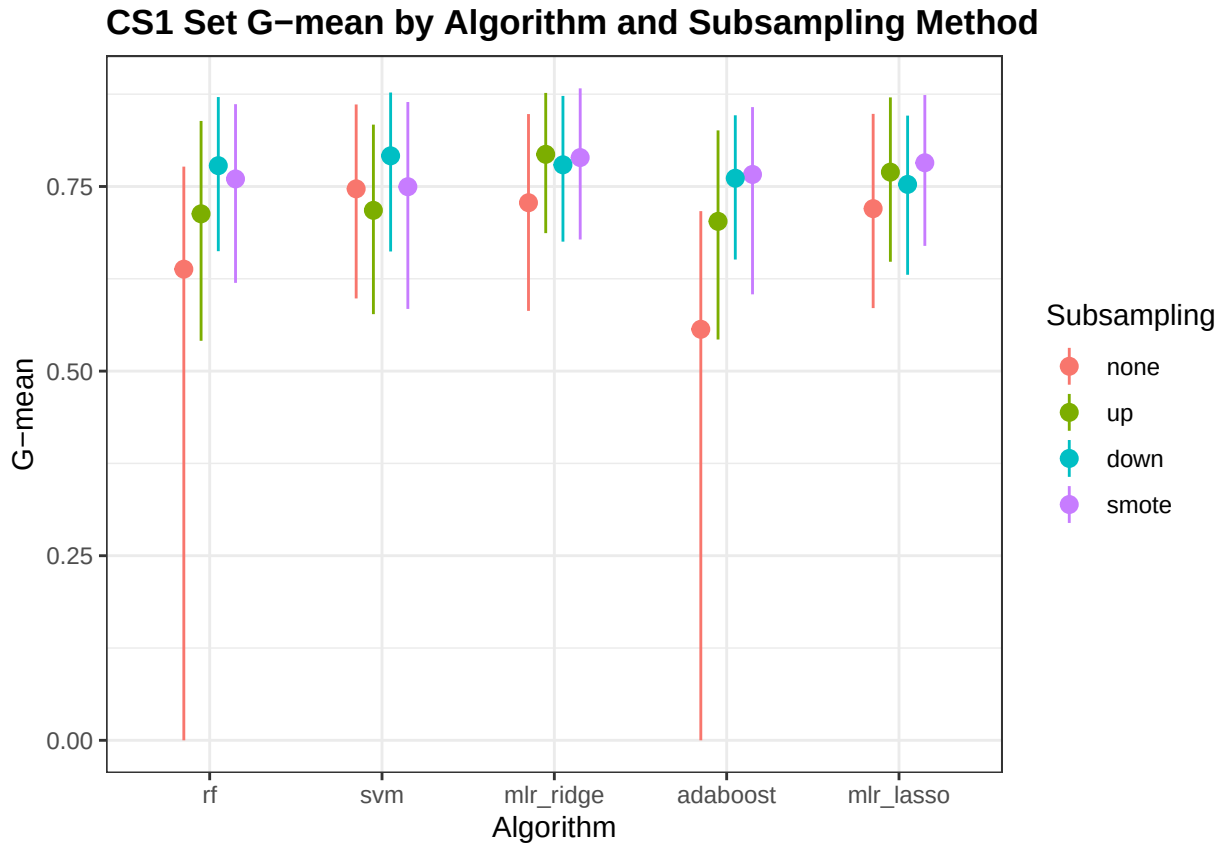

Figure 4.31: CS1 Set G-mean

Table 4.30: CS1 Set Class-Specific G-mean by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.889	0.889	0.889	0.884	0.879
none	ENOC	0.854	0.868	0.845	0.831	0.851
none	HGSC	0.892	0.896	0.905	0.871	0.896
none	LGSC	0.577	0.812	0.707	0.408	0.707
none	MUC	0.756	0.791	0.836	0.745	0.816
up	CCOC	0.893	0.886	0.889	0.886	0.868
up	ENOC	0.863	0.848	0.847	0.848	0.836
up	HGSC	0.909	0.895	0.906	0.898	0.899
up	LGSC	0.707	0.793	0.88	0.703	0.873
up	MUC	0.812	0.745	0.866	0.808	0.853
down	CCOC	0.896	0.889	0.897	0.89	0.883
down	ENOC	0.847	0.859	0.848	0.817	0.823
down	HGSC	0.873	0.866	0.862	0.861	0.848
down	LGSC	0.871	0.865	0.877	0.866	0.851
down	MUC	0.854	0.902	0.856	0.855	0.837
smote	CCOC	0.895	0.892	0.889	0.893	0.887
smote	ENOC	0.87	0.868	0.856	0.856	0.853
smote	HGSC	**0.913**	0.897	0.902	0.904	0.89
smote	LGSC	0.804	0.812	0.861	0.812	0.857
smote	MUC	0.832	0.791	0.857	0.82	0.861

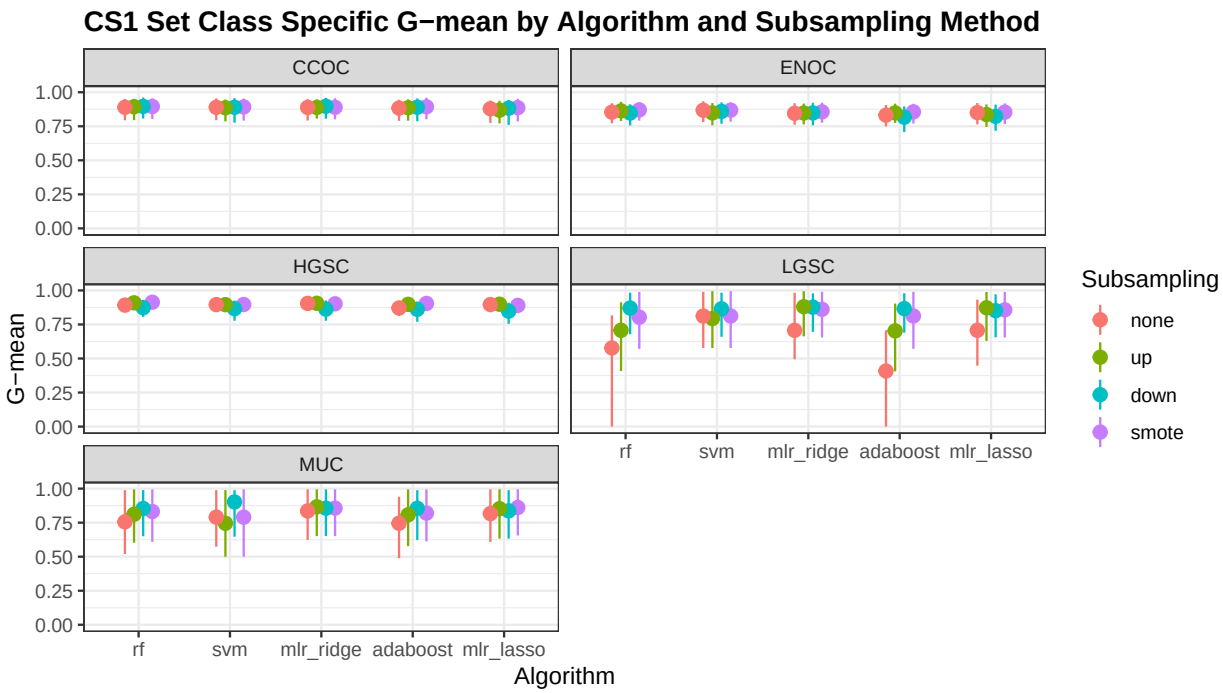

Figure 4.32: CS1 Set Class-Specific G-mean

4.4 CS2 Set

Table 4.31: CS2 Set Accuracy by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.921	0.925	**0.933**	0.905	0.927
up	0.925	0.923	0.918	0.93	0.916
down	0.855	0.838	0.812	0.839	0.811
smote	0.925	0.92	0.909	0.921	0.896

4.4.1 Accuracy

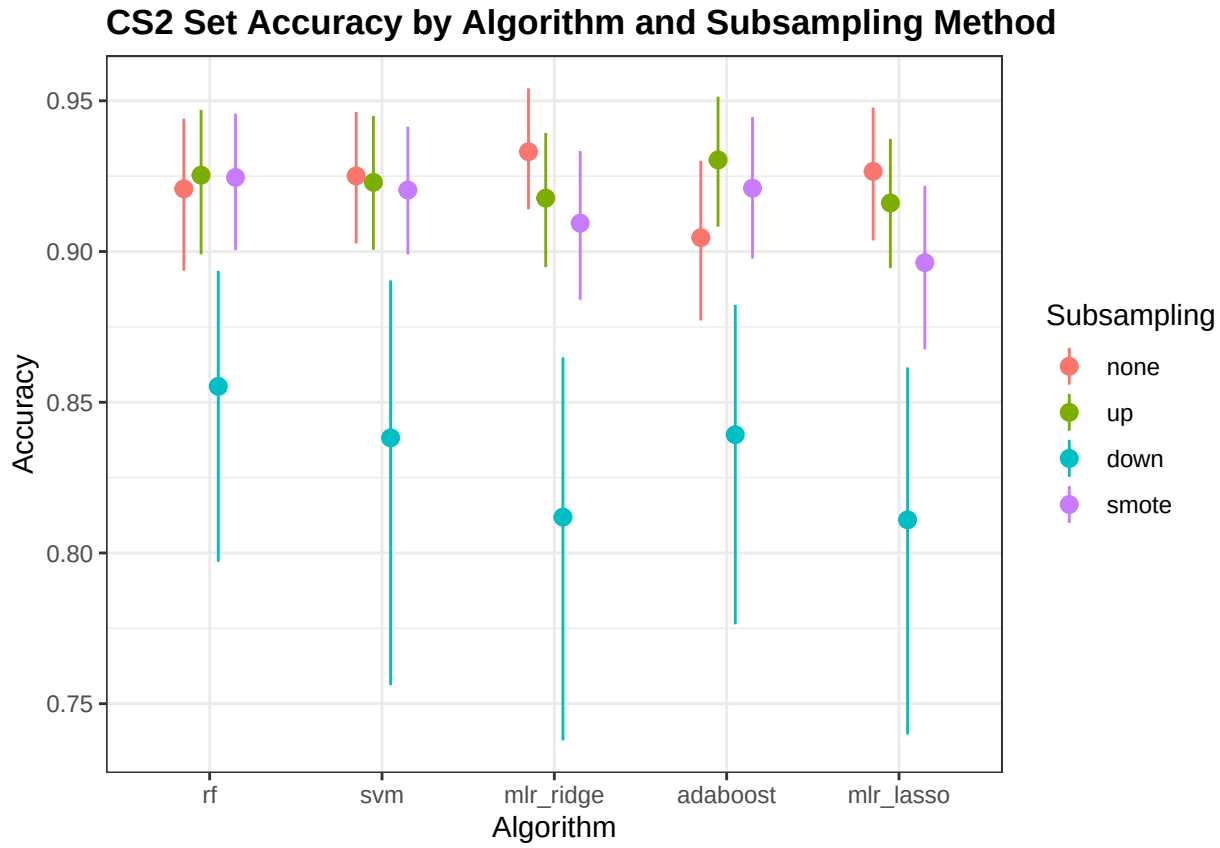

Figure 4.33: CS2 Set Accuracy

Table 4.32: CS2 Set Class-Specific Accuracy by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.984	0.981	**0.987**	0.981	0.983
none	ENOC	0.973	0.978	0.977	0.965	0.974
none	HGSC	0.927	0.935	0.946	0.908	0.943
none	LGSC	0.977	0.977	0.977	0.977	0.976
none	MUC	0.981	0.978	0.983	0.98	0.98
up	CCOC	0.986	0.98	**0.987**	0.986	0.986
up	ENOC	0.977	0.979	0.967	0.98	0.966
up	HGSC	0.93	0.93	0.935	0.94	0.936
up	LGSC	0.977	0.98	0.971	0.978	0.972
up	MUC	0.982	0.978	0.975	0.98	0.976
down	CCOC	0.981	0.953	0.976	0.978	0.971
down	ENOC	0.958	0.954	0.952	0.957	0.939
down	HGSC	0.876	0.862	0.84	0.865	0.839
down	LGSC	0.952	0.956	0.921	0.94	0.922
down	MUC	0.95	0.96	0.945	0.946	0.961
smote	CCOC	0.985	0.979	0.986	0.984	0.981
smote	ENOC	0.975	0.977	0.963	0.974	0.957
smote	HGSC	0.939	0.932	0.928	0.939	0.919
smote	LGSC	0.98	0.98	0.97	0.98	0.964
smote	MUC	0.972	0.975	0.974	0.968	0.973

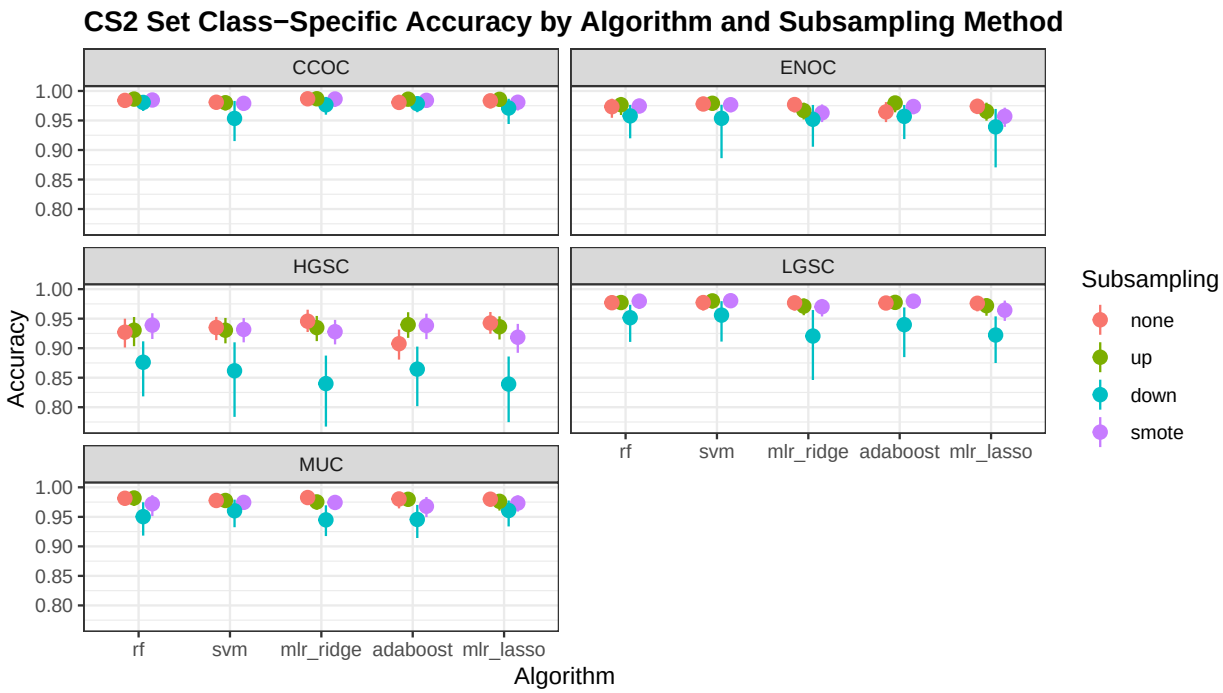

Figure 4.34: CS2 Set Class-Specific Accuracy

Table 4.33: CS2 Set Macro-Averaged F1-Score by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.718	0.762	0.752	0.736	0.74
up	0.719	0.751	0.773	0.74	0.754
down	0.699	0.668	0.652	0.675	0.645
smote	**0.782**	0.755	0.762	0.769	0.732

4.4.2 F1-Score

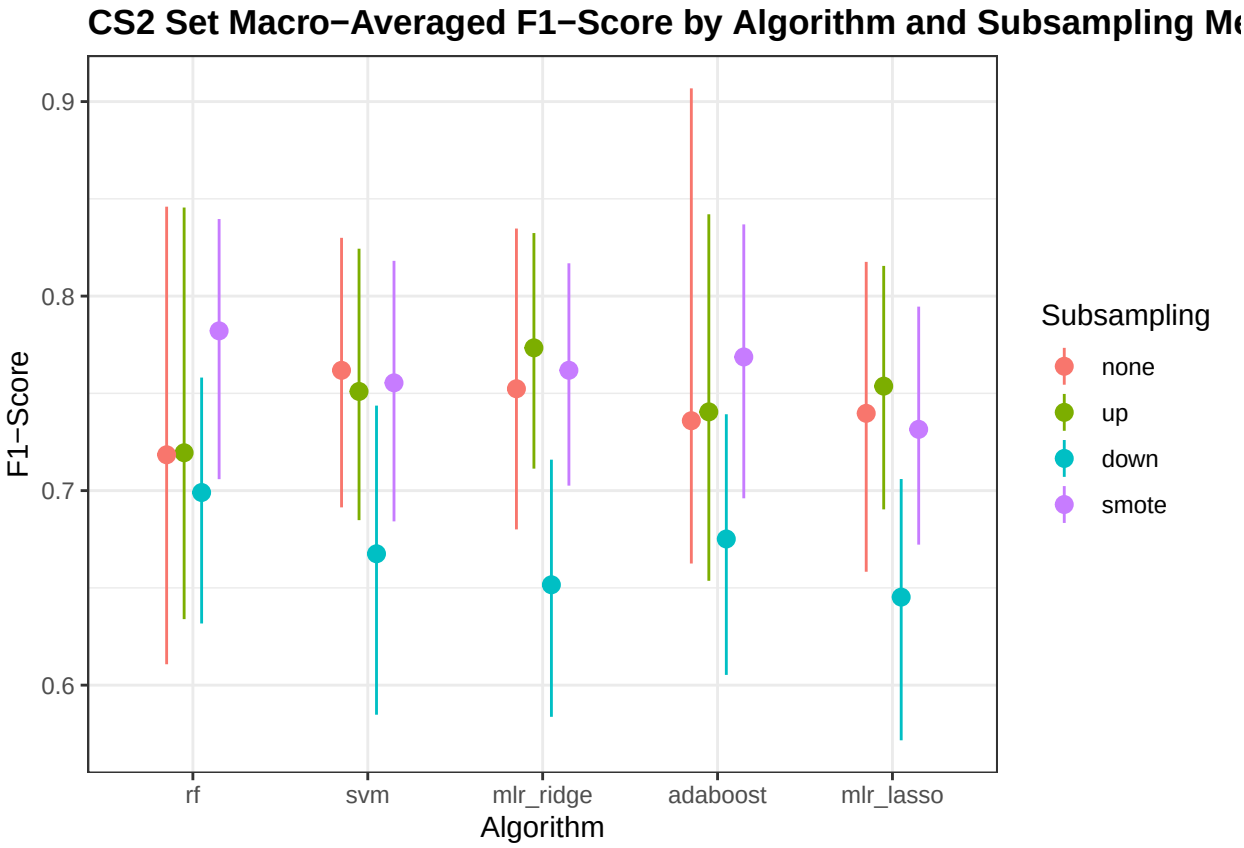

Figure 4.35: CS2 Set F1-Score

Table 4.34: CS2 Set Class-Specific F1-Score by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.889	0.857	0.905	0.857	0.878
none	ENOC	0.5	0.667	0.636	0.222	0.609
none	HGSC	0.955	0.96	**0.966**	0.945	0.964
none	LGSC	0.222	0.5	0.364	0.268	0.4
none	MUC	0.87	0.842	0.875	0.851	0.863
up	CCOC	0.894	0.842	0.909	0.9	0.896
up	ENOC	0.571	0.667	0.606	0.667	0.581
up	HGSC	0.958	0.957	0.958	0.963	0.959
up	LGSC	0.25	0.5	0.571	0.308	0.526
up	MUC	0.864	0.833	0.84	0.86	0.837
down	CCOC	0.875	0.739	0.84	0.863	0.809
down	ENOC	0.563	0.519	0.515	0.545	0.452
down	HGSC	0.916	0.907	0.888	0.908	0.89
down	LGSC	0.435	0.444	0.341	0.375	0.34
down	MUC	0.72	0.742	0.696	0.7	0.75
smote	CCOC	0.9	0.839	0.903	0.898	0.872
smote	ENOC	0.667	0.632	0.581	0.645	0.533
smote	HGSC	0.961	0.958	0.953	0.961	0.947
smote	LGSC	0.556	0.545	0.545	0.556	0.5
smote	MUC	0.821	0.818	0.833	0.8	0.824

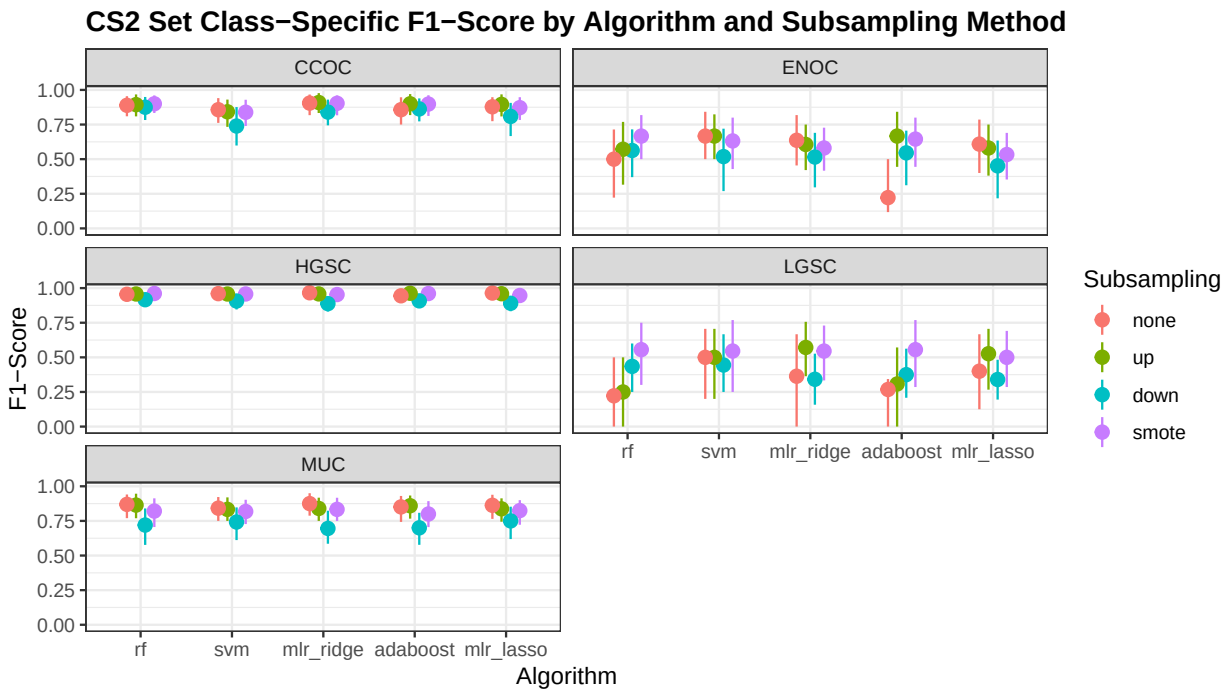

Figure 4.36: CS2 Set Class-Specific F1-Score

Table 4.35: CS2 Set Kappa by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.75	0.774	**0.803**	0.678	0.788
up	0.763	0.76	0.787	0.791	0.775
down	0.67	0.629	0.601	0.642	0.594
smote	0.798	0.762	0.77	0.788	0.736

4.4.3 Kappa

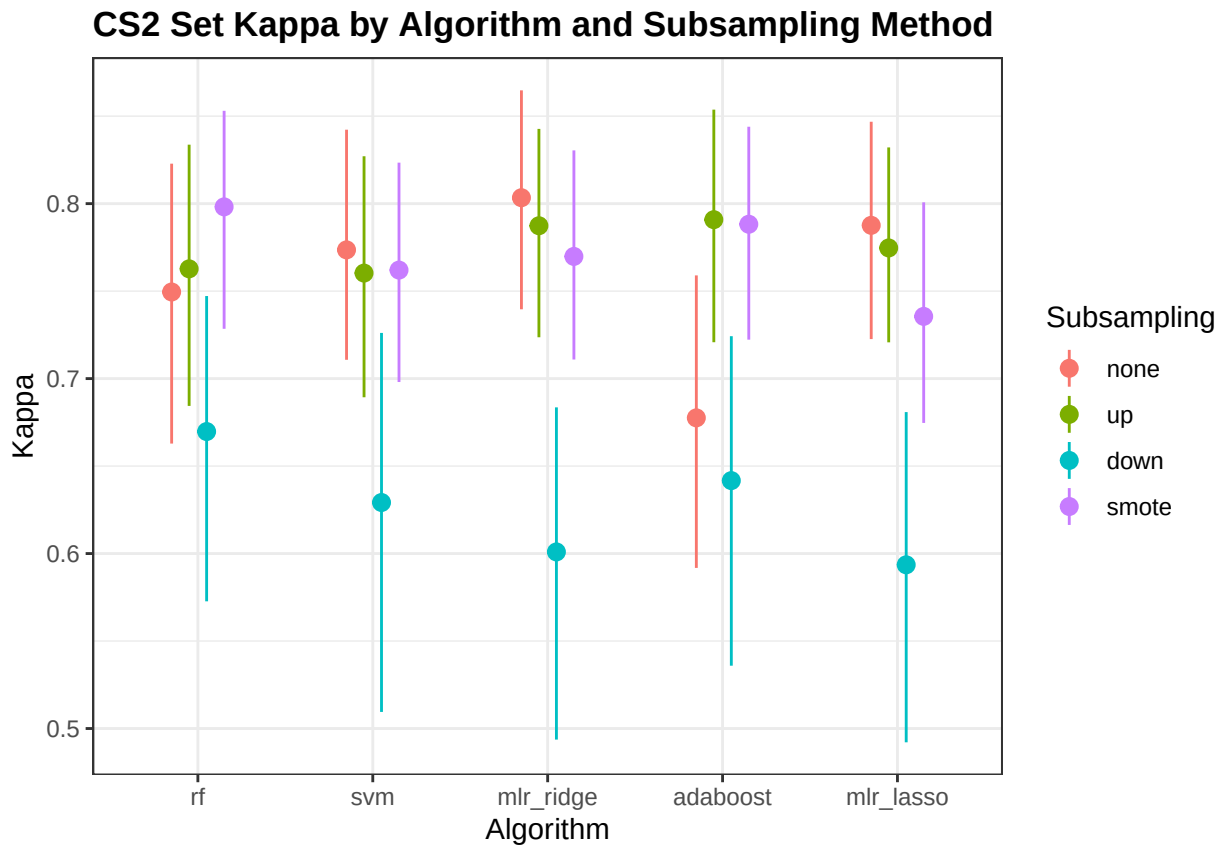

Figure 4.37: CS2 Set Kappa

Table 4.36: CS2 Set Class-Specific Kappa by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.88	0.849	0.897	0.846	0.869
none	ENOC	0.484	0.661	0.623	0.162	0.595
none	HGSC	0.754	0.786	0.828	0.669	0.82
none	LGSC	0.173	0.486	0.353	0	0.386
none	MUC	0.859	0.831	0.865	0.84	0.85
up	CCOC	0.885	0.831	**0.902**	0.893	0.888
up	ENOC	0.559	0.656	0.588	0.655	0.559
up	HGSC	0.761	0.769	0.813	0.804	0.812
up	LGSC	0.214	0.488	0.558	0.282	0.512
up	MUC	0.854	0.823	0.826	0.848	0.825
down	CCOC	0.864	0.714	0.826	0.85	0.793
down	ENOC	0.542	0.494	0.49	0.528	0.423
down	HGSC	0.68	0.645	0.614	0.657	0.605
down	LGSC	0.418	0.425	0.313	0.352	0.312
down	MUC	0.693	0.72	0.665	0.673	0.73
smote	CCOC	0.893	0.83	0.897	0.889	0.863
smote	ENOC	0.655	0.62	0.559	0.627	0.511
smote	HGSC	0.82	0.778	0.796	0.817	0.771
smote	LGSC	0.542	0.523	0.531	0.542	0.482
smote	MUC	0.805	0.805	0.82	0.783	0.809

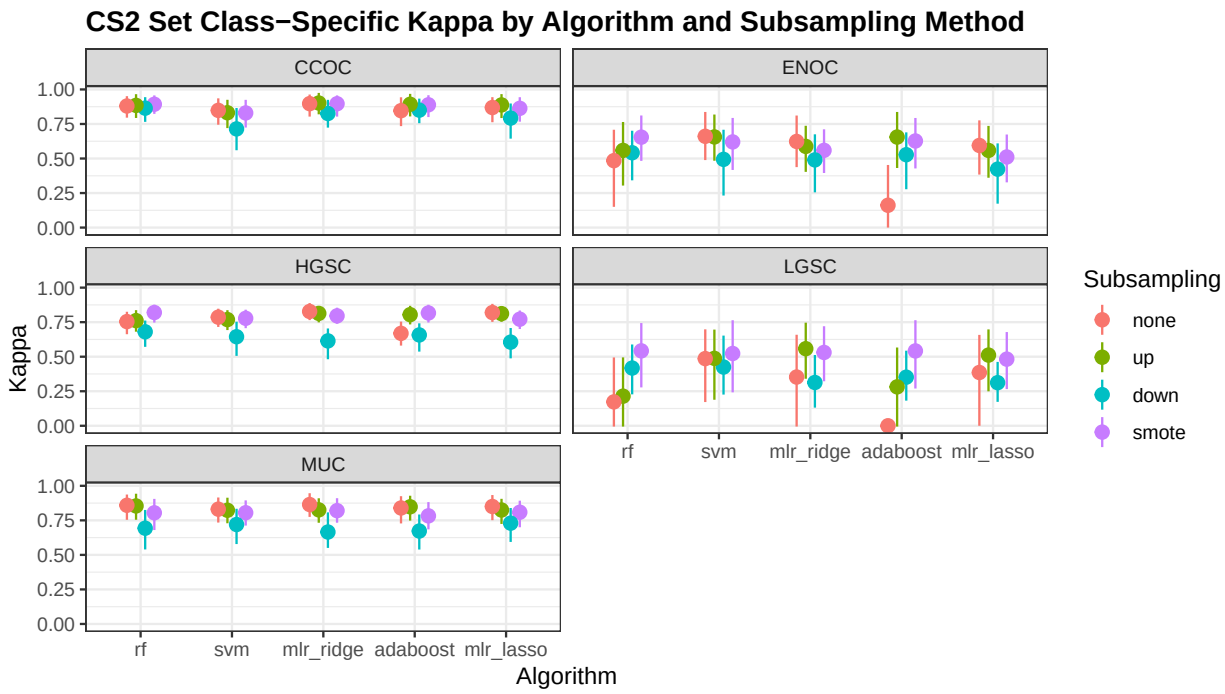

Figure 4.38: CS2 Set Class-Specific Kappa

Table 4.37: CS2 Set G-mean by Algorithm and Subsampling Method

sampling	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	0.401	0.683	0.645	0	0.654
up	0.508	0.642	**0.835**	0.589	0.766
down	0.829	0.792	0.801	0.804	0.786
smote	0.776	0.677	0.825	0.763	0.801

4.4.4 G-mean

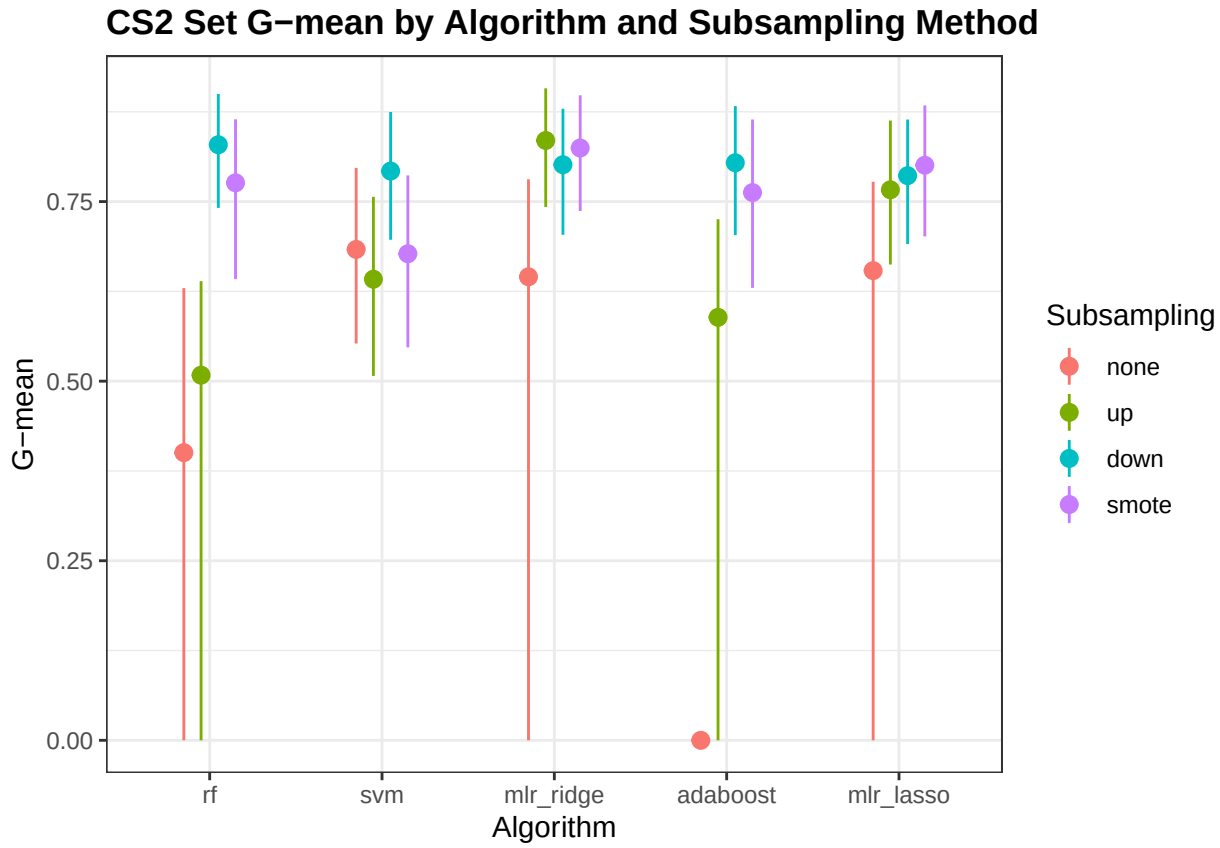

Figure 4.39: CS2 Set G-mean

Table 4.38: CS2 Set Class-Specific G-mean by Algorithm and Subsampling Method

sampling	histotype	rf	svm	mlr_ridge	adaboost	mlr_lasso
none	CCOC	0.909	0.888	0.934	0.872	0.924
none	ENOC	0.577	0.764	0.737	0.302	0.73
none	HGSC	0.824	0.861	0.888	0.754	0.892
none	LGSC	0.316	0.667	0.534	0	0.575
none	MUC	0.918	0.884	0.932	0.887	0.932
up	CCOC	0.909	0.858	**0.965**	0.929	0.941
up	ENOC	0.632	0.737	0.81	0.73	0.78
up	HGSC	0.826	0.836	0.933	0.871	0.919
up	LGSC	0.354	0.628	0.902	0.446	0.807
up	MUC	0.905	0.87	0.935	0.928	0.922
down	CCOC	0.959	0.931	0.927	0.948	0.917
down	ENOC	0.832	0.809	0.796	0.803	0.78
down	HGSC	0.901	0.884	0.878	0.891	0.872
down	LGSC	0.892	0.886	0.884	0.872	0.891
down	MUC	0.916	0.882	0.92	0.917	0.901
smote	CCOC	0.956	0.877	0.961	0.954	0.941
smote	ENOC	0.812	0.734	0.808	0.796	0.793
smote	HGSC	0.919	0.857	0.928	0.918	0.919
smote	LGSC	0.744	0.703	0.887	0.749	0.861
smote	MUC	0.93	0.88	0.93	0.926	0.919

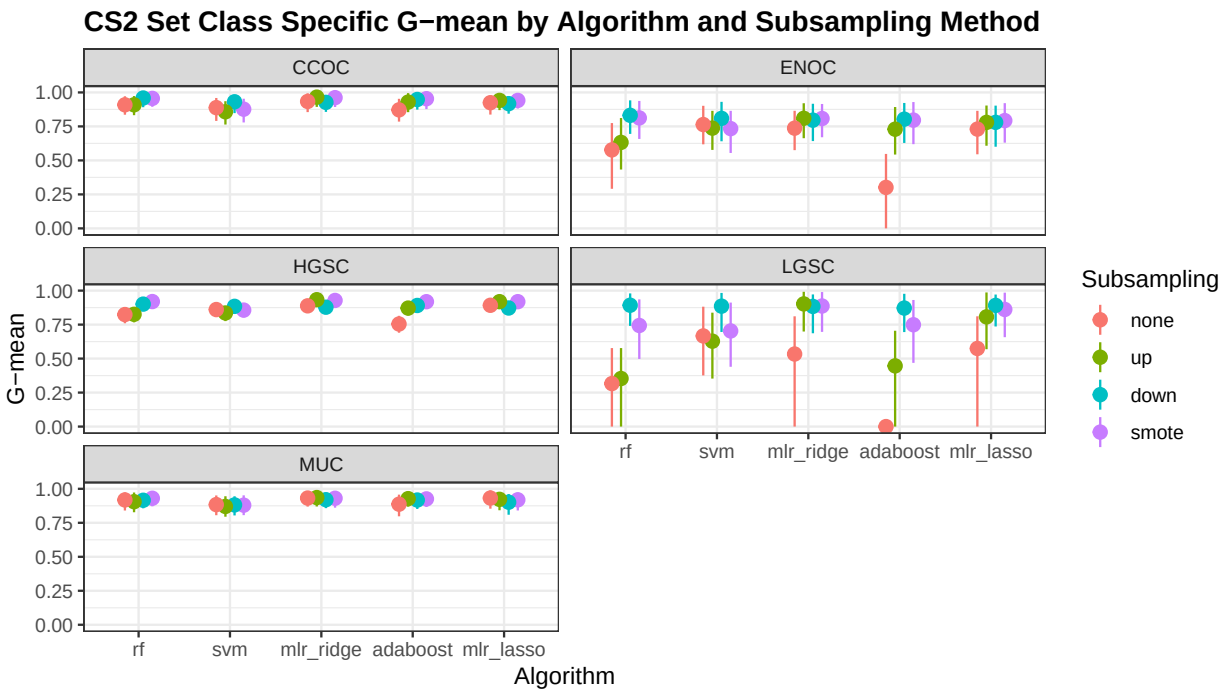

Figure 4.40: CS2 Set Class-Specific G-mean

Table 4.39: SMOTE Kappa by Algorithm and Dataset

dataset	rf	svm	mlr_ridge	adaboost	mlr_lasso
Training	**0.83**	0.81	0.766	0.809	0.758
CS1	0.763	0.759	0.755	0.751	0.736
CS2	0.798	0.762	0.77	0.788	0.736

4.5 SMOTE Kappa Summary

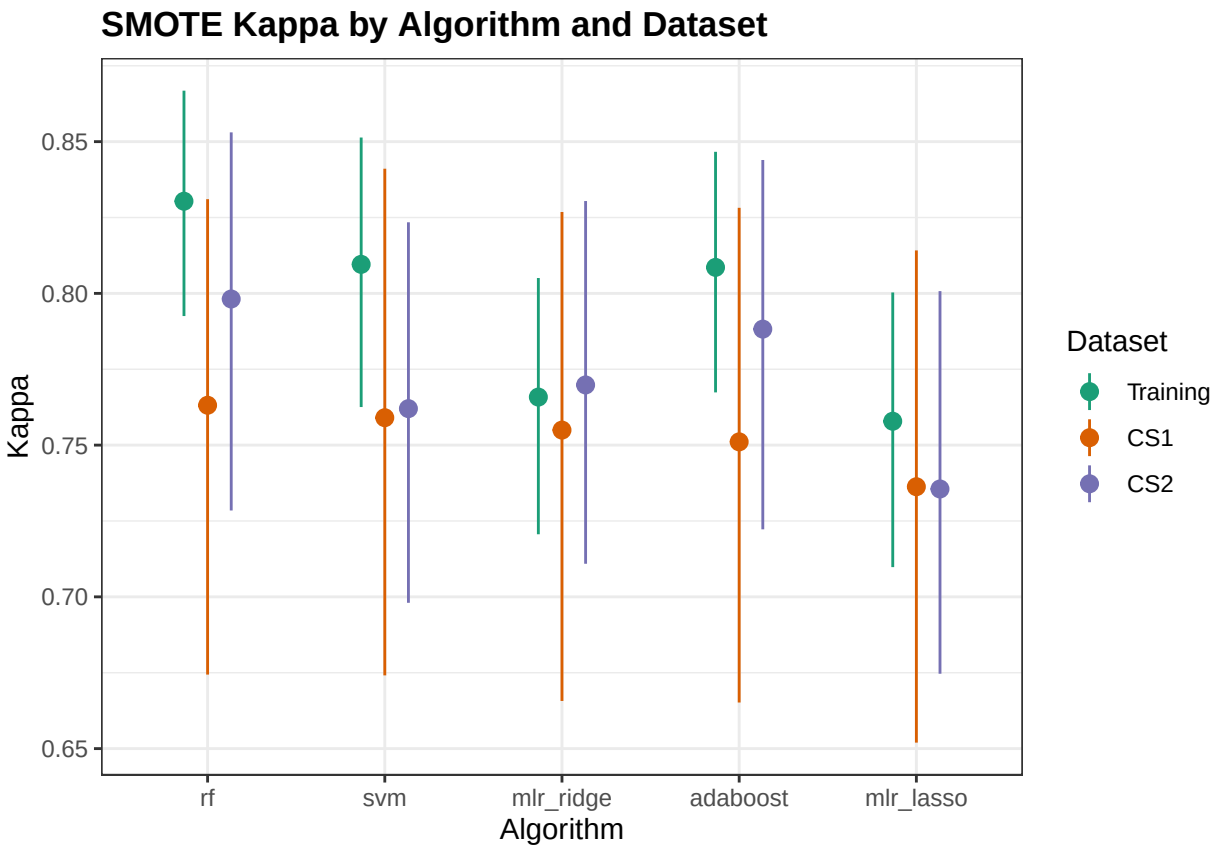

Figure 4.41: SMOTE Kappa by Algorithm and Dataset

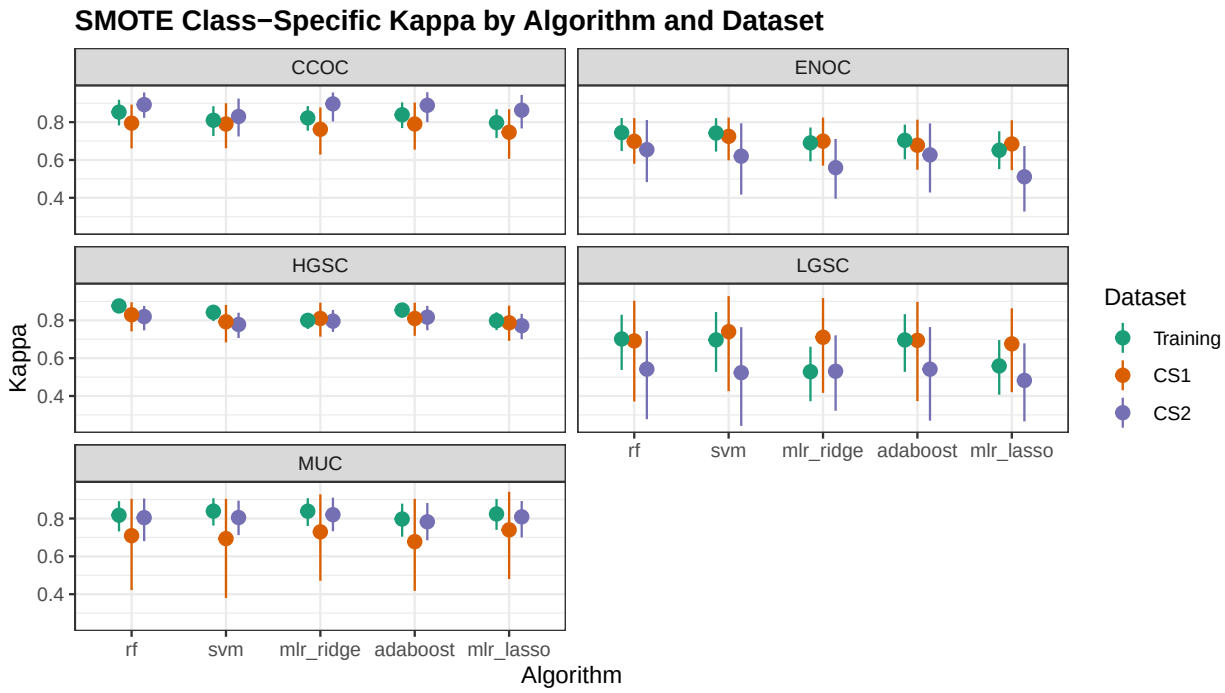

Figure 4.42: SMOTE Class-Specific Kappa by Algorithm and Dataset

4.6 Overlap with SPOT

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.

4.7 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
2S-rf-none	0.92	0.831	0.974	0.88	0.889
sequential-none	0.904	0.972	0.971	0.933	0.893
2S-rf-smote	0.914	0.833	0.972	0.877	0.892
sequential-smote	0.905	0.971	0.972	0.929	0.906
sequential-up	0.892	0.97	0.971	0.929	0.889
2S-svm-none	0.892	0.844	0.972	0.917	0.879
2S-rf-up	0.919	0.831	0.973	0.875	0.889
2S-svm-up	0.891	0.847	0.973	0.917	0.877
2S-adaboost-up	0.905	0.824	0.971	0.88	0.876
sequential-down	0.904	0.971	0.963	0.933	0.893
2S-adaboost-smote	0.901	0.818	0.97	0.882	0.871
2S-mlr_ridge-none	0.903	0.833	0.966	0.886	0.904
2S-mlr_ridge-up	0.899	0.833	0.96	0.867	0.899
2S-mlr_lasso-none	0.899	0.822	0.965	0.857	0.885
2S-mlr_ridge-down	0.892	0.827	0.957	0.846	0.889
2S-mlr_lasso-up	0.896	0.82	0.958	0.857	0.881
2S-svm-smote	0.886	0.845	0.971	0.909	0.875
2S-rf-down	0.907	0.812	0.962	0.868	0.877
2S-adaboost-none	0.911	0.818	0.97	0.88	0.867
mlr_ridge-none	0.853	0.757	0.967	0.353	0.872
2S-mlr_ridge-smote	0.898	0.833	0.964	0.857	0.897
2S-adaboost-down	0.897	0.806	0.963	0.87	0.867
rf-up	0.864	0.767	0.969	0.522	0.852
rf-none	0.866	0.724	0.967	0.375	0.852
2S-mlr_lasso-smote	0.897	0.822	0.963	0.857	0.877
2S-svm-down	0.875	0.825	0.96	0.909	0.862
2S-mlr_lasso-down	0.895	0.812	0.953	0.839	0.857
mlr_lasso-none	0.843	0.74	0.968	0.471	0.849
rf-smote	0.862	0.759	0.974	0.71	0.829
adaboost-up	0.865	0.754	0.974	0.69	0.839
adaboost-smote	0.85	0.721	0.969	0.706	0.81
svm-none	0.848	0.786	0.97	0.667	0.857
mlr_ridge-smote	0.835	0.71	0.953	0.545	0.847
mlr_ridge-up	0.833	0.667	0.944	0.524	0.841
svm-up	0.822	0.75	0.969	0.714	0.846
mlr_lasso-up	0.778	0.638	0.954	0.611	0.808
svm-smote	0.824	0.759	0.968	0.706	0.847
mlr_lasso-smote	0.813	0.675	0.954	0.571	0.833
adaboost-none	0.846	0.681	0.959	0.2	0.827
rf-down	0.841	0.641	0.93	0.487	0.765
svm-down	0.79	0.629	0.928	0.526	0.762
mlr_ridge-down	0.821	0.627	0.917	0.413	0.811
adaboost-down	0.833	0.615	0.927	0.481	0.741
mlr_lasso-down	0.792	0.605	0.907	0.368	0.769

Showing 1 to 44 of 44 entries Previous 1 Next

Table 4.40: Class-specific F1-scores on Confirmation Sets

method	HGSC	CCOC	ENOC	LGSC	MUC
2S-rf-none	0.915	0.904	0.772	0.419	0.679
sequential-none	0.905	0.917	0.917	0.485	0.711
2S-rf-smote	0.908	0.878	0.750	0.383	0.691
sequential-smote	0.907	0.923	0.916	0.514	0.679
rf-none	0.893	0.887	0.486	0.133	0.745

The 44 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.

4.8 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 4 methods are:

- **2S-rf-none**: 2-step method using random forest algorithm with no subsampling
- **sequential-none**: sequential algorithm with no subsampling. The sequence of models and algorithms used are:
  - HGSC vs. non-HGSC using SVM
  - MUC vs. non-MUC using ridge regression
  - CCOC vs. non-CCOC using random forest
  - ENOC vs. LGSC using SVM
- **2S-rf-smote**: 2-step method using random forest algorithm with SMOTE subsampling
- **sequential-smote**: sequential algorithm with SMOTE subsampling. The sequence of models and algorithms used are:
  - HGSC vs. non-HGSC using random forest
  - CCOC vs. non-CCOC using random forest
  - MUC vs. non-MUC using SVM
  - ENOC vs. LGSC using random forest

As a comparison we also show the F1-scores from the **rf-none** to see how the best methods improve from it.

4.8.1 Confirmation Set

In the confirmation set, **2S-rf-none** improves drastically in LGSC classification compard to **rf-none**, with moderate improvement in ENOC, and minor improvement in HGSC and CCOC. There is a decrease in MUC performance. **sequential-none** improves on **2S-rf-none** in all classes except for marginal decrease in HGSC performance.

4.8.2 Validation Set

Similarly in the validation set, **2S-rf-none** improves drastically in LGSC classification compared to **rf-none**, with large improvement in ENOC, and minor improvement in HGSC and CCOC. There is a decrease in MUC performance. **sequential-none** improves on **2S-rf-none** in all classes except for a small decrease in CCOC performance and same performance for LGSC.

Table 4.41: Class-specific F1-scores on Validation Sets

method	HGSC	CCOC	ENOC	LGSC	MUC
2S-rf-none	0.931	0.920	0.823	0.750	0.571
sequential-none	0.943	0.862	0.964	0.750	0.725
2S-rf-smote	0.917	0.920	0.788	0.732	0.538
sequential-smote	0.926	0.902	0.953	0.698	0.617
rf-none	0.920	0.839	0.540	0.174	0.686