# Ovarian Cancer Histotypes: Report of Statistical Findings

Derek Chiu

2023-03-07

# 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 = 273
  - Genes = 72

The final CS1 training set has 254 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 | 123 | 645 | 1645 |
| non-HGSC | 166 | 220 | 585 |

Table 3.2: All CodeSet Major Reviewed Histotypes

| Reviewed Histotype | CS1 | CS2 | CS3 | CS1 % | CS2 % | CS3 % |
|---|---|---|---|---|---|---|
| CCOC | 48 | 61 | 175 | 17.8 | 7.4 | 8.1 |
| ENOC | 60 | 32 | 232 | 22.2 | 3.9 | 10.7 |
| HGSC | 123 | 645 | 1645 | 45.6 | 78.6 | 76.1 |
| LGSC | 20 | 21 | 40 | 7.4 | 2.6 | 1.9 |
| MUC | 19 | 62 | 69 | 7.0 | 7.6 | 3.2 |

Table 3.3: All CodeSet Reviewed Histotypes

| Reviewed Histotype | CS1 | CS2 | CS3 |
|---|---|---|---|
| CARCINOMA-NOS | 0 | 1 | 23 |
| CCOC | 48 | 61 | 175 |
| CTRL | 0 | 12 | 0 |
| ENOC | 60 | 32 | 232 |
| HGSC | 123 | 645 | 1645 |
| 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 | 29 |
| 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 | 123 |
| 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 | 645 |
| 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 | 175 |
| CS3 | ENOC | 232 |
| CS3 | HGSC | 1645 |
| CS3 | LGSC | 40 |
| CS3 | MBOT | 3 |
| CS3 | MIXED (ENOC/CCOC) | 1 |
| CS3 | MIXED (ENOC/LGSC) | 1 |
| CS3 | MIXED (HGSC/CCOC) | 1 |
| CS3 | MMMT | 29 |
| 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 | 57 | 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 | 55 | 58 | 70 |
| 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 = 881
  - Genes = 365

The final CS2 training set has 821 samples on 136 (common) genes after normalization and keeping only the major histotypes of interest.

## 3.3 Common Samples

## 3.4 Histotypes in Classifier Data

Table 3.11: Distinct Common Samples Histotype Distribution

| revHist | CS1 | CS2 | CS3 |
|---|---|---|---|
| CCOC | 3 | 3 | 3 |
| ENOC | 3 | 3 | 3 |
| HGSC | 53 | 53 | 53 |
| 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 | 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 | 119 (48%) | 624 (80%) | 475 (94%) | 1218 (79%) |
| | CCOC | 44 (18%) | 54 (7%) | 8 (2%) | 106 (7%) |
| | ENOC | 55 (22%) | 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 (%) | 247 (16%) | 783 (51%) | 506 (33%) | 1536 (100%) |

Table 3.18: Histotype Distribution by CodeSet/Datasets

| Variable | Levels | CS1 All | CS2 All | Confirmation | Validation |
|---|---|---|---|---|---|
| Histotype | HGSC | 122 (46%) | 644 (79%) | 422 (66%) | 676 (74%) |
| | CCOC | 47 (18%) | 60 (7%) | 75 (12%) | 81 (9%) |
| | ENOC | 58 (22%) | 30 (4%) | 106 (16%) | 108 (12%) |
| | MUC | 18 (7%) | 61 (7%) | 27 (4%) | 26 (3%) |
| | LGSC | 18 (7%) | 20 (2%) | 13 (2%) | 18 (2%) |
| Total | N (%) | 263 (10%) | 815 (31%) | 643 (24%) | 909 (35%) |

# 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

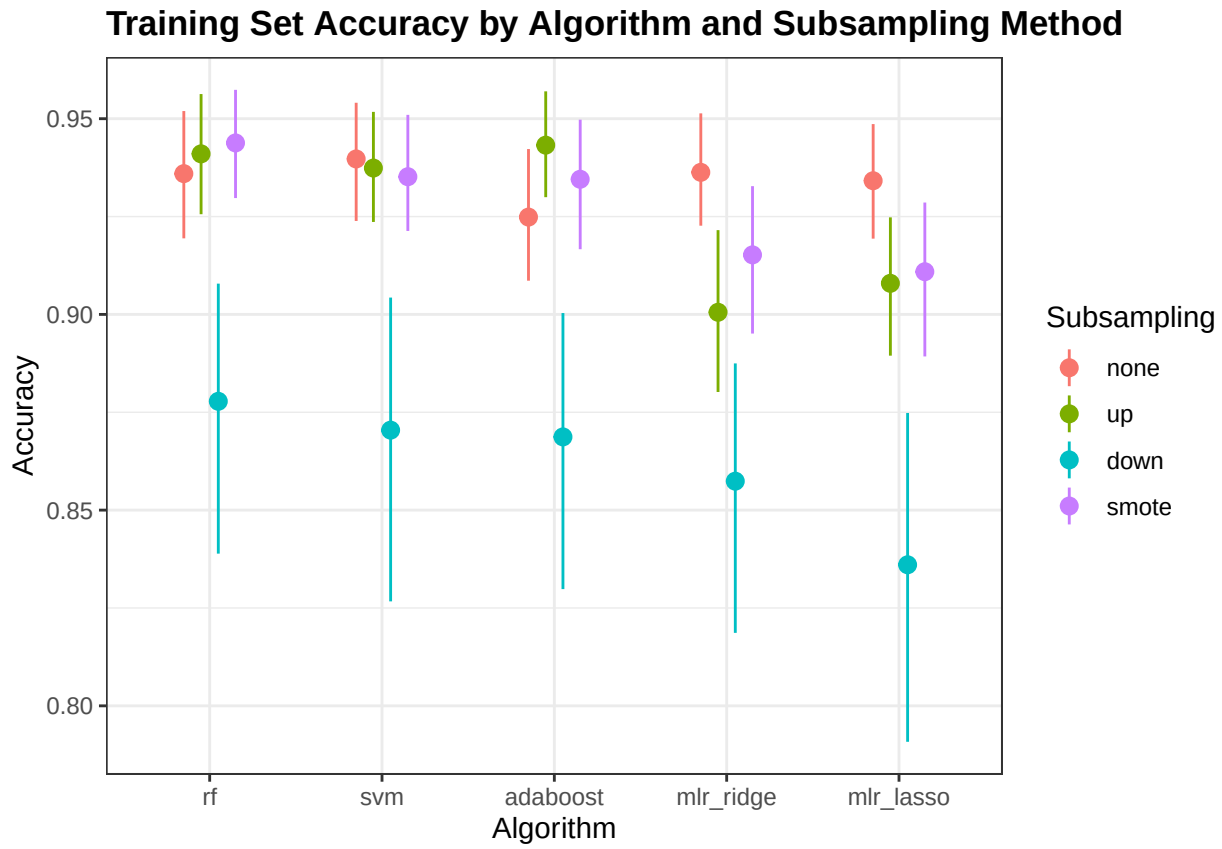


Figure 4.1: Training Set Accuracy

Table 4.1: Training Set Accuracy by Algorithm and Subsampling Method

| sampling | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|---|---|---|---|---|---|
| none | 0.936 | 0.94 | 0.925 | 0.936 | 0.934 |
| up | 0.941 | 0.937 | 0.943 | 0.901 | 0.908 |
| down | 0.878 | 0.87 | 0.869 | 0.857 | 0.836 |
| smote | **0.944** | 0.935 | 0.935 | 0.915 | 0.911 |

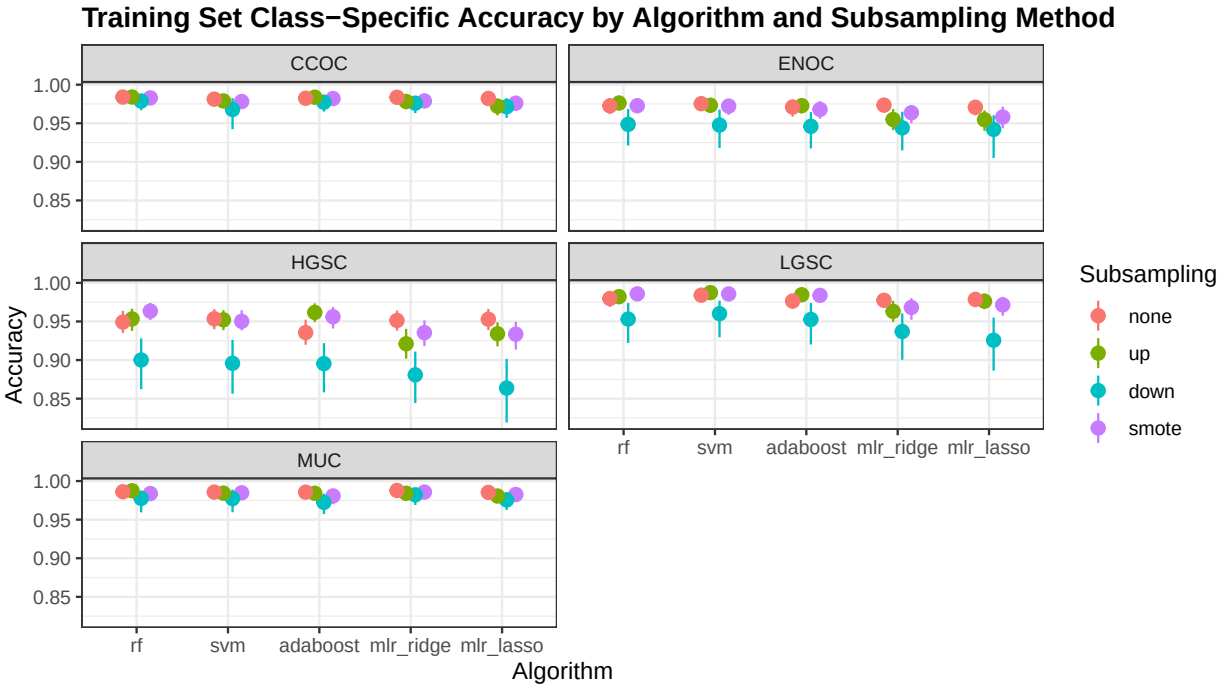


Figure 4.2: Training Set Class-Specific Accuracy

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

| sampling | histotype | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|---|---|---|---|---|---|---|
| none | CCOC | 0.984 | 0.981 | 0.982 | 0.984 | 0.982 |
| none | ENOC | 0.973 | 0.975 | 0.971 | 0.973 | 0.971 |
| none | HGSC | 0.949 | 0.953 | 0.936 | 0.951 | 0.953 |
| none | LGSC | 0.98 | 0.984 | 0.976 | 0.977 | 0.979 |
| none | MUC | 0.986 | 0.986 | 0.986 | **0.988** | 0.985 |
| up | CCOC | 0.984 | 0.979 | 0.984 | 0.978 | 0.972 |
| up | ENOC | 0.976 | 0.973 | 0.973 | 0.955 | 0.955 |
| up | HGSC | 0.953 | 0.952 | 0.962 | 0.921 | 0.934 |
| up | LGSC | 0.982 | 0.987 | 0.985 | 0.963 | 0.976 |
| up | MUC | 0.987 | 0.984 | 0.984 | 0.984 | 0.981 |
| down | CCOC | 0.979 | 0.968 | 0.977 | 0.976 | 0.971 |
| down | ENOC | 0.949 | 0.948 | 0.946 | 0.944 | 0.942 |
| down | HGSC | 0.9 | 0.896 | 0.895 | 0.881 | 0.864 |
| down | LGSC | 0.953 | 0.96 | 0.953 | 0.937 | 0.926 |
| down | MUC | 0.977 | 0.977 | 0.972 | 0.982 | 0.976 |
| smote | CCOC | 0.983 | 0.978 | 0.982 | 0.979 | 0.976 |
| smote | ENOC | 0.973 | 0.972 | 0.968 | 0.964 | 0.958 |
| smote | HGSC | 0.963 | 0.95 | 0.956 | 0.935 | 0.934 |
| smote | LGSC | 0.986 | 0.986 | 0.984 | 0.968 | 0.971 |
| smote | MUC | 0.984 | 0.985 | 0.981 | 0.986 | 0.983 |

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

| sampling | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|---|---|---|---|---|---|
| none | 0.762 | 0.823 | 0.728 | 0.769 | 0.781 |
| up | 0.802 | 0.823 | 0.826 | 0.775 | 0.767 |
| down | 0.742 | 0.73 | 0.725 | 0.727 | 0.694 |
| smote | **0.835** | 0.823 | 0.818 | 0.791 | 0.779 |

### 4.1.2 F1-Score

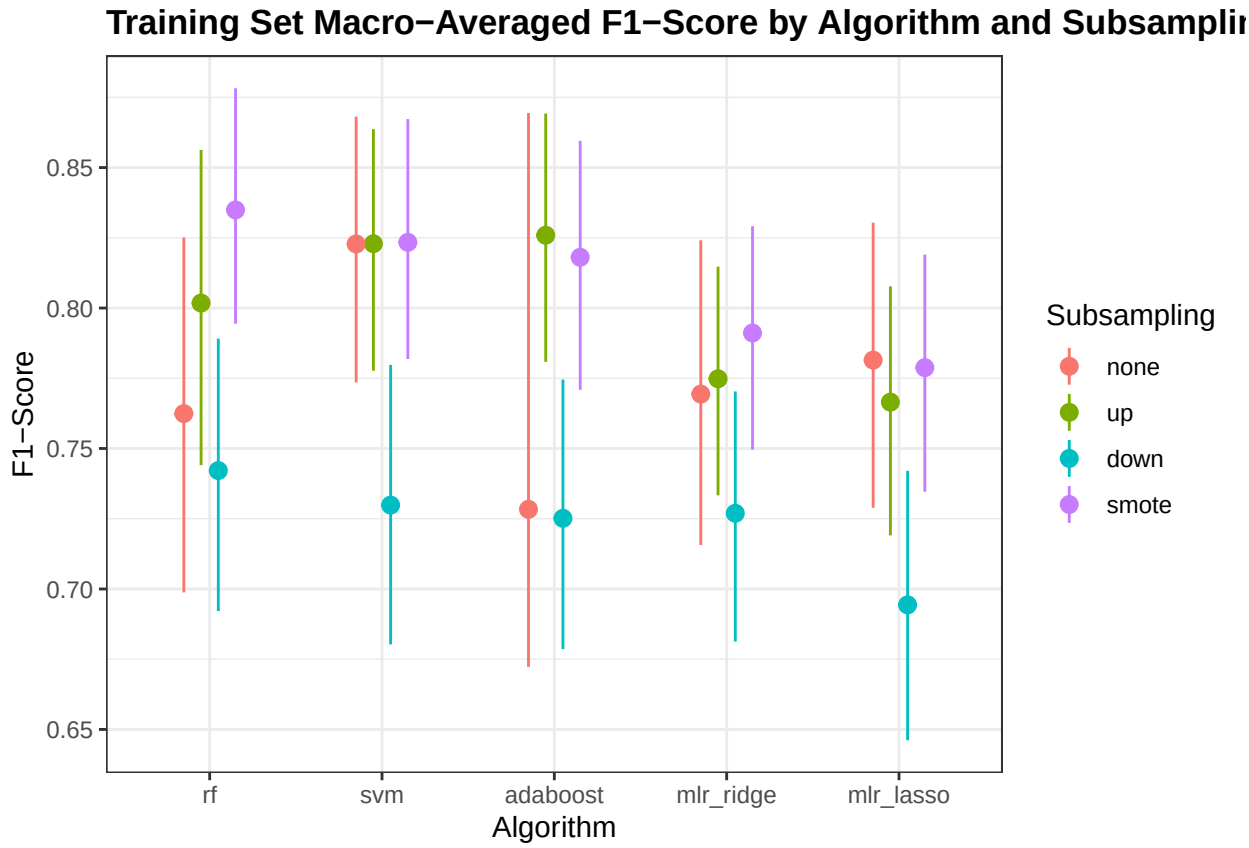


Figure 4.3: Training Set F1-Score

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

| sampling | histotype | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|---|---|---|---|---|---|---|
| none | CCOC | 0.875 | 0.857 | 0.862 | 0.871 | 0.865 |
| none | ENOC | 0.733 | 0.776 | 0.702 | 0.758 | 0.734 |
| none | HGSC | 0.969 | 0.971 | 0.961 | 0.97 | 0.971 |
| none | LGSC | 0.375 | 0.667 | 0.167 | 0.37 | 0.483 |
| none | MUC | 0.873 | 0.857 | 0.857 | 0.885 | 0.862 |
| up | CCOC | 0.873 | 0.836 | 0.873 | 0.846 | 0.8 |
| up | ENOC | 0.776 | 0.754 | 0.765 | 0.675 | 0.646 |
| up | HGSC | 0.971 | 0.971 | 0.976 | 0.948 | 0.958 |
| up | LGSC | 0.526 | 0.72 | 0.667 | 0.545 | 0.615 |
| up | MUC | 0.877 | 0.846 | 0.857 | 0.862 | 0.821 |
| down | CCOC | 0.849 | 0.781 | 0.837 | 0.829 | 0.805 |
| down | ENOC | 0.645 | 0.639 | 0.617 | 0.625 | 0.602 |
| down | HGSC | 0.934 | 0.931 | 0.929 | 0.92 | 0.907 |
| down | LGSC | 0.481 | 0.519 | 0.473 | 0.418 | 0.377 |
| down | MUC | 0.8 | 0.8 | 0.771 | 0.841 | 0.785 |
| smote | CCOC | 0.873 | 0.831 | 0.865 | 0.85 | 0.83 |
| smote | ENOC | 0.769 | 0.762 | 0.732 | 0.716 | 0.676 |
| smote | HGSC | **0.977** | 0.969 | 0.972 | 0.959 | 0.957 |
| smote | LGSC | 0.714 | 0.71 | 0.692 | 0.571 | 0.588 |
| smote | MUC | 0.849 | 0.851 | 0.833 | 0.868 | 0.847 |

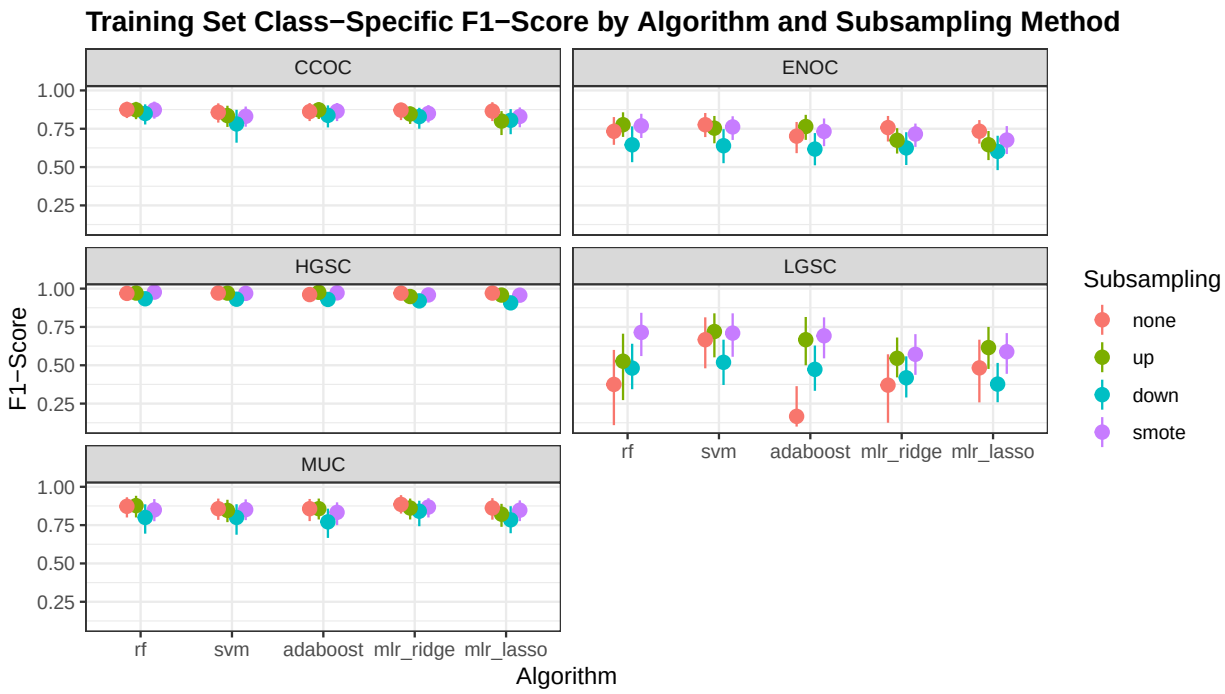


Figure 4.4: Training Set Class-Specific F1-Score

Table 4.5: Training Set Kappa by Algorithm and Subsampling Method

| sampling | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|---|---|---|---|---|---|
| none | 0.804 | 0.824 | 0.76 | 0.81 | 0.81 |
| up | 0.821 | 0.814 | 0.838 | 0.754 | 0.757 |
| down | 0.71 | 0.694 | 0.691 | 0.677 | 0.638 |
| smote | **0.843** | 0.816 | 0.821 | 0.782 | 0.769 |

### 4.1.3 Kappa

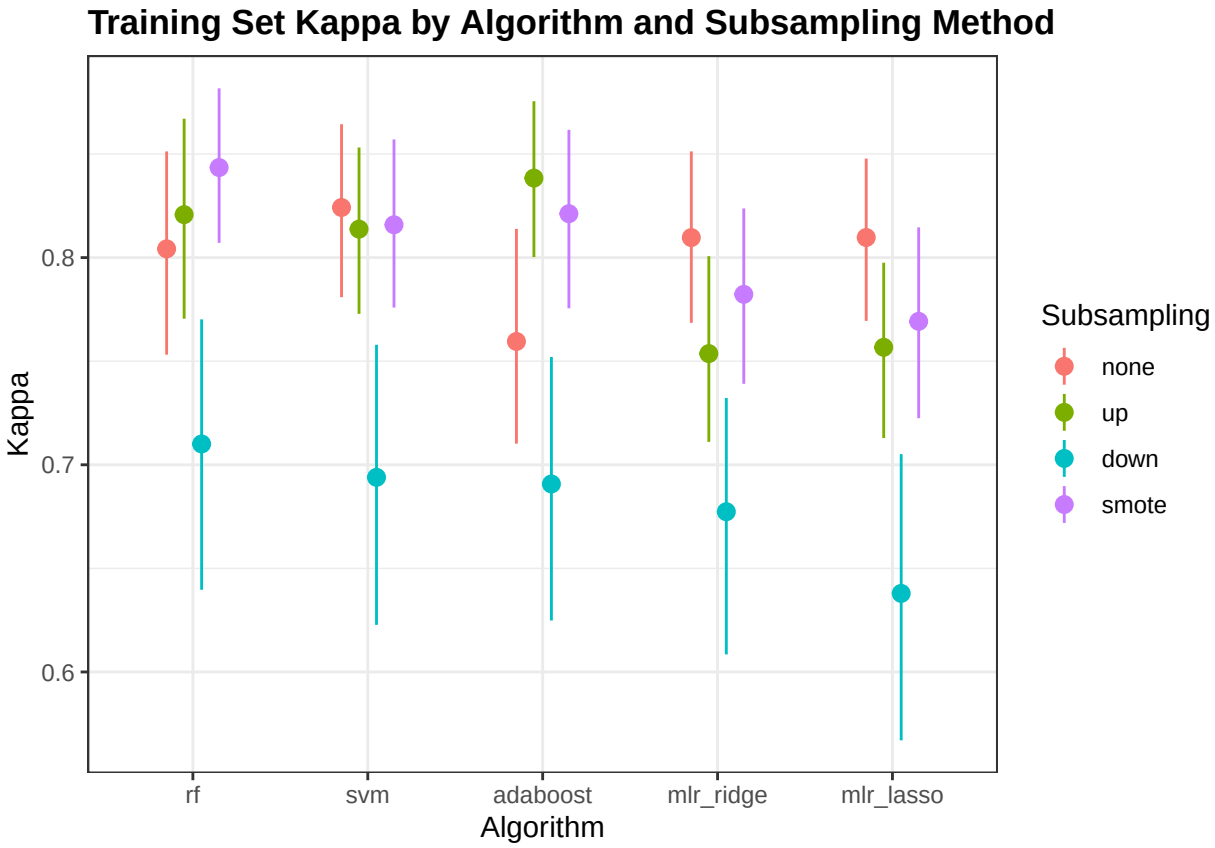


Figure 4.5: Training Set Kappa

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

| sampling | histotype | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|---|---|---|---|---|---|---|
| none | CCOC | 0.868 | 0.846 | 0.853 | 0.862 | 0.855 |
| none | ENOC | 0.719 | 0.763 | 0.686 | 0.743 | 0.718 |
| none | HGSC | 0.832 | 0.854 | 0.781 | 0.843 | 0.853 |
| none | LGSC | 0.368 | 0.659 | 0.122 | 0.357 | 0.469 |
| none | MUC | 0.866 | 0.849 | 0.849 | 0.879 | 0.853 |
| up | CCOC | 0.865 | 0.824 | 0.864 | 0.835 | 0.784 |
| up | ENOC | 0.764 | 0.738 | 0.75 | 0.649 | 0.623 |
| up | HGSC | 0.846 | 0.845 | 0.881 | 0.784 | 0.806 |
| up | LGSC | 0.518 | 0.713 | 0.66 | 0.529 | 0.602 |
| up | MUC | 0.871 | 0.839 | 0.849 | 0.853 | 0.812 |
| down | CCOC | 0.838 | 0.763 | 0.824 | 0.817 | 0.79 |
| down | ENOC | 0.617 | 0.614 | 0.589 | 0.598 | 0.571 |
| down | HGSC | 0.734 | 0.723 | 0.724 | 0.695 | 0.659 |
| down | LGSC | 0.46 | 0.502 | 0.452 | 0.394 | 0.35 |
| down | MUC | 0.787 | 0.79 | 0.756 | 0.83 | 0.772 |
| smote | CCOC | 0.865 | 0.82 | 0.855 | 0.838 | 0.817 |
| smote | ENOC | 0.756 | 0.746 | 0.714 | 0.695 | 0.655 |
| smote | HGSC | **0.89** | 0.846 | 0.867 | 0.818 | 0.81 |
| smote | LGSC | 0.707 | 0.702 | 0.685 | 0.558 | 0.576 |
| smote | MUC | 0.839 | 0.844 | 0.822 | 0.861 | 0.839 |

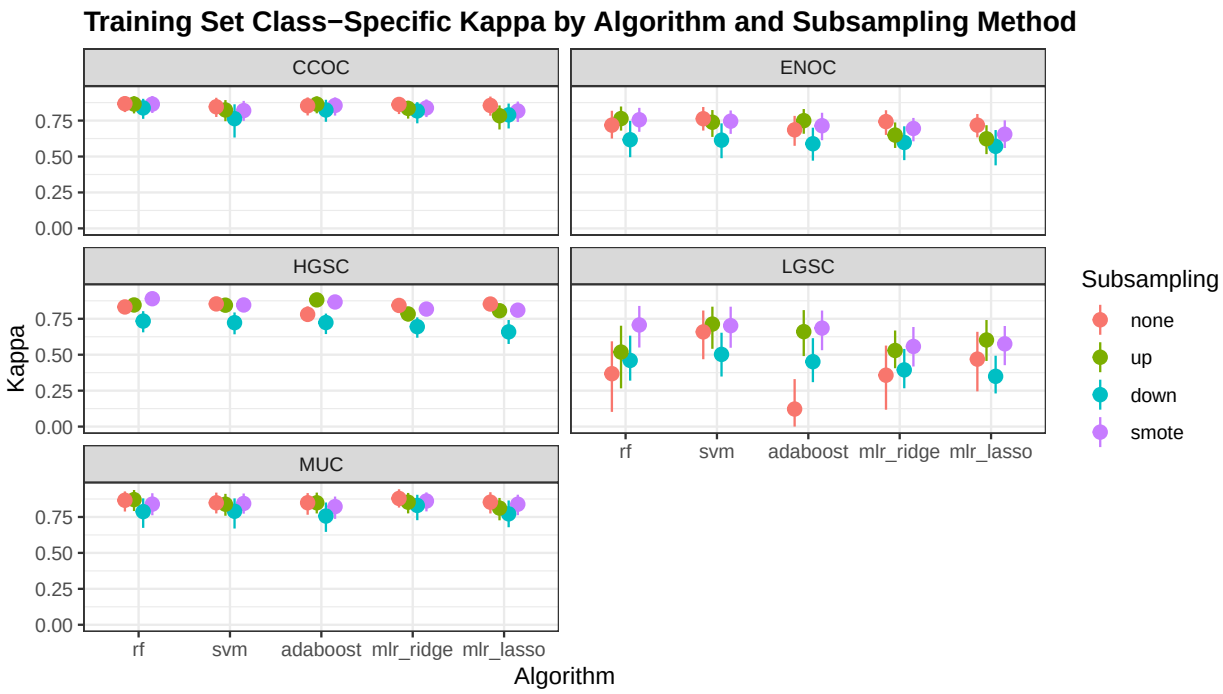


Figure 4.6: Training Set Class-Specific Kappa

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

| sampling | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|---|---|---|---|---|---|
| none | 0.638 | 0.785 | 0.474 | 0.667 | 0.721 |
| up | 0.714 | 0.77 | 0.802 | **0.87** | 0.803 |
| down | 0.859 | 0.848 | 0.845 | 0.861 | 0.837 |
| smote | 0.838 | 0.803 | 0.835 | 0.864 | 0.839 |

### 4.1.4 G-mean

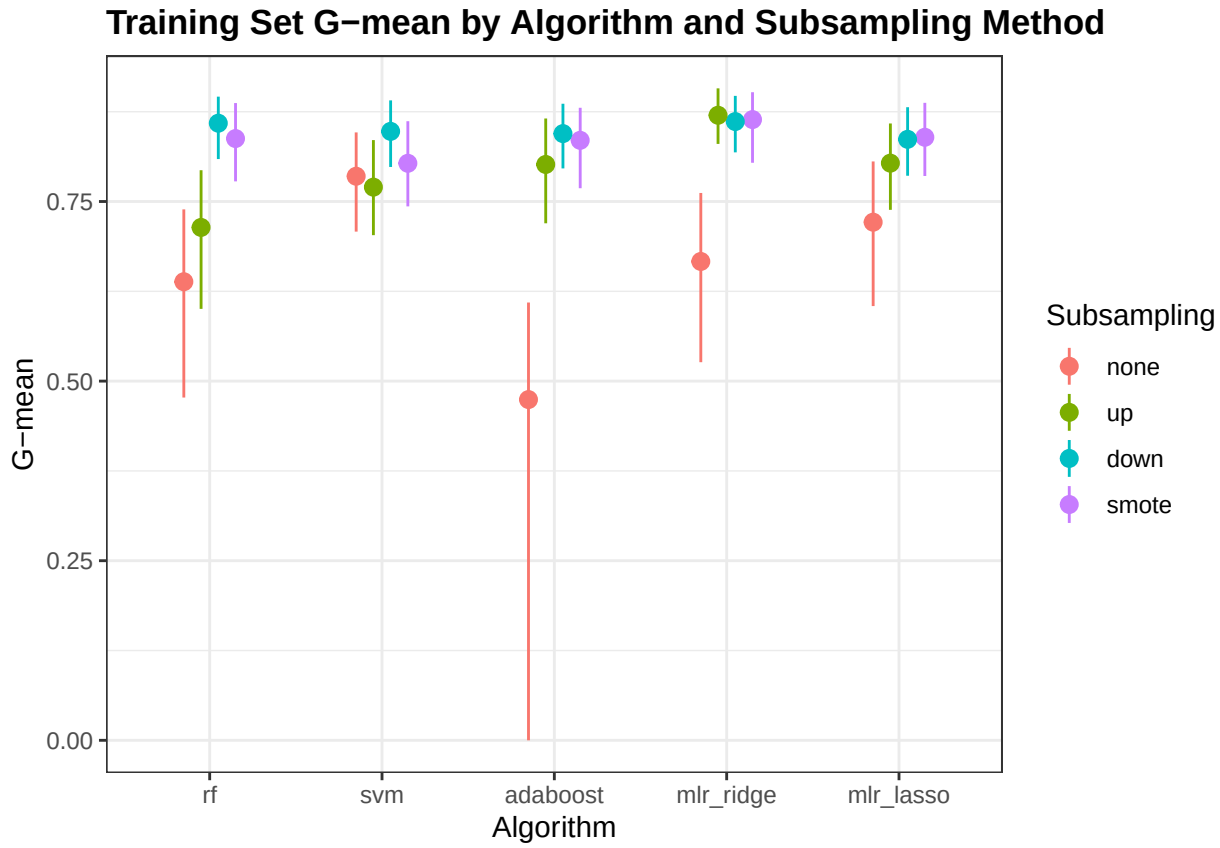


Figure 4.7: Training Set G-mean

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

| sampling | histotype | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|---|---|---|---|---|---|---|
| none | CCOC | 0.909 | 0.9 | 0.893 | 0.906 | 0.914 |
| none | ENOC | 0.804 | 0.853 | 0.763 | 0.839 | 0.829 |
| none | HGSC | 0.877 | 0.907 | 0.835 | 0.894 | 0.908 |
| none | LGSC | 0.485 | 0.794 | 0.258 | 0.516 | 0.639 |
| none | MUC | 0.924 | 0.901 | 0.899 | 0.931 | 0.921 |
| up | CCOC | 0.903 | 0.875 | 0.913 | 0.928 | 0.895 |
| up | ENOC | 0.834 | 0.832 | 0.866 | 0.878 | 0.83 |
| up | HGSC | 0.889 | 0.893 | 0.932 | 0.935 | 0.921 |
| up | LGSC | 0.623 | 0.814 | 0.78 | 0.935 | 0.865 |
| up | MUC | 0.923 | 0.886 | 0.932 | 0.94 | 0.905 |
| down | CCOC | 0.927 | 0.91 | 0.921 | 0.921 | 0.917 |
| down | ENOC | 0.878 | 0.877 | 0.858 | 0.877 | 0.858 |
| down | HGSC | 0.922 | 0.916 | 0.917 | 0.91 | 0.899 |
| down | LGSC | 0.915 | 0.92 | 0.912 | 0.931 | 0.912 |
| down | MUC | 0.926 | 0.916 | 0.926 | 0.932 | 0.913 |
| smote | CCOC | 0.929 | 0.891 | 0.921 | 0.927 | 0.921 |
| smote | ENOC | 0.875 | 0.867 | 0.861 | 0.874 | 0.857 |
| smote | HGSC | **0.947** | 0.912 | 0.943 | 0.939 | 0.933 |
| smote | LGSC | 0.847 | 0.848 | 0.864 | 0.922 | 0.895 |
| smote | MUC | 0.93 | 0.9 | 0.93 | 0.935 | 0.925 |

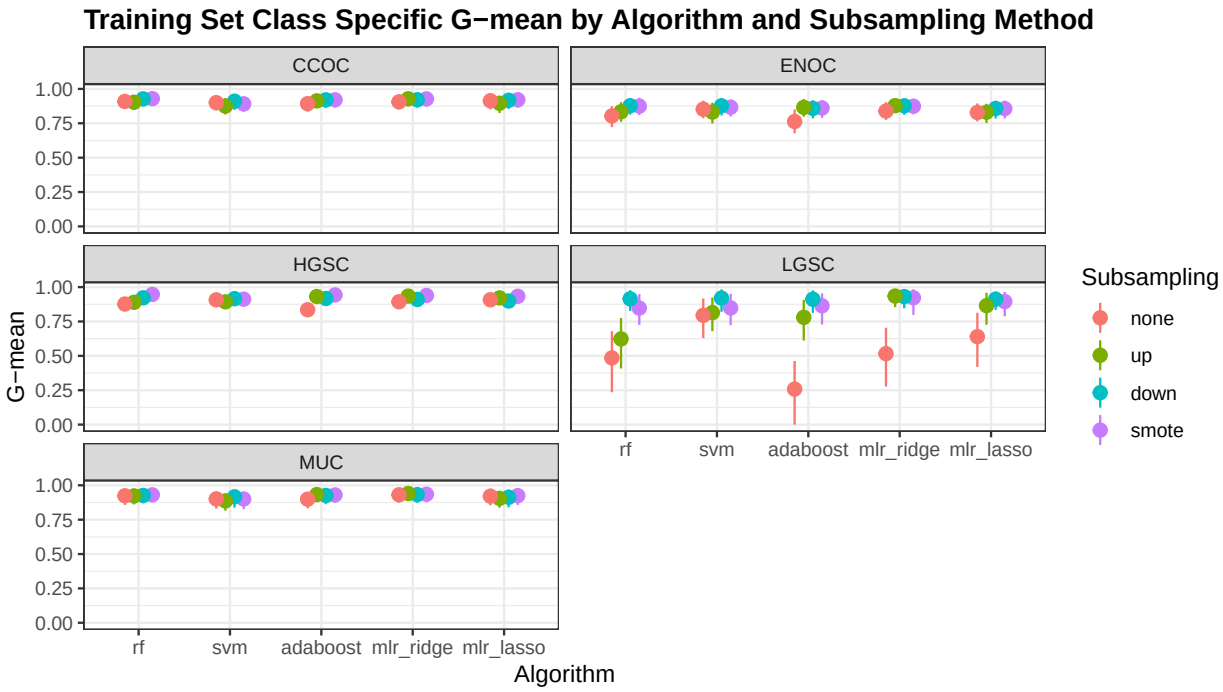


Figure 4.8: Training Set Class-Specific G-mean

## 4.2 Two-Step Training Set

### 4.2.1 Accuracy

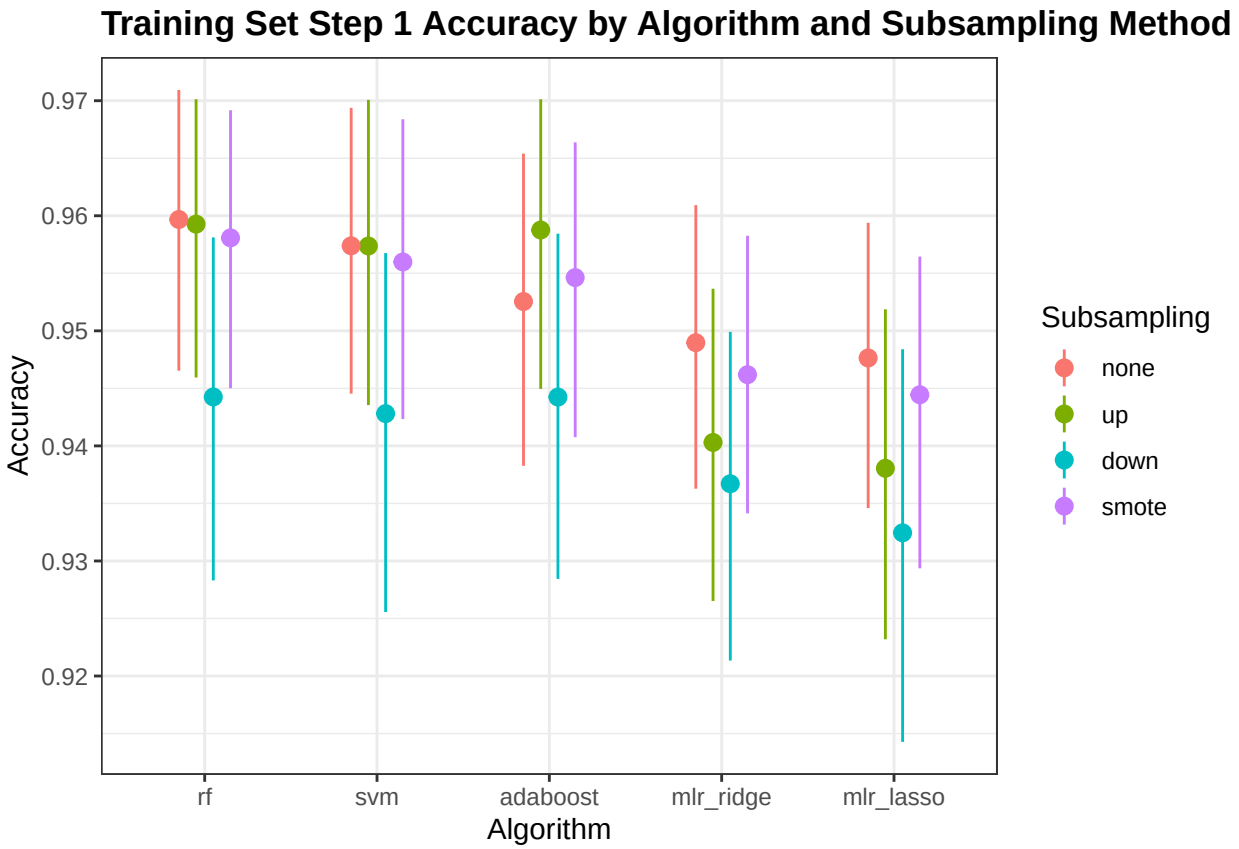


Figure 4.9: Training Set Step 1 Accuracy

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

| sampling | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|---|---|---|---|---|---|
| none | **0.96** | 0.957 | 0.953 | 0.949 | 0.948 |
| up | 0.959 | 0.957 | 0.959 | 0.94 | 0.938 |
| down | 0.944 | 0.943 | 0.944 | 0.937 | 0.932 |
| smote | 0.958 | 0.956 | 0.955 | 0.946 | 0.944 |

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

| sampling | rf | svm | mlr_ridge | adaboost | mlr_lasso |
|---|---|---|---|---|---|
| none | 0.884 | 0.882 | 0.881 | 0.876 | 0.872 |
| up | **0.886** | 0.881 | 0.878 | 0.874 | 0.87 |
| down | 0.868 | 0.868 | 0.868 | 0.861 | 0.854 |
| smote | 0.88 | 0.877 | 0.875 | 0.871 | 0.87 |

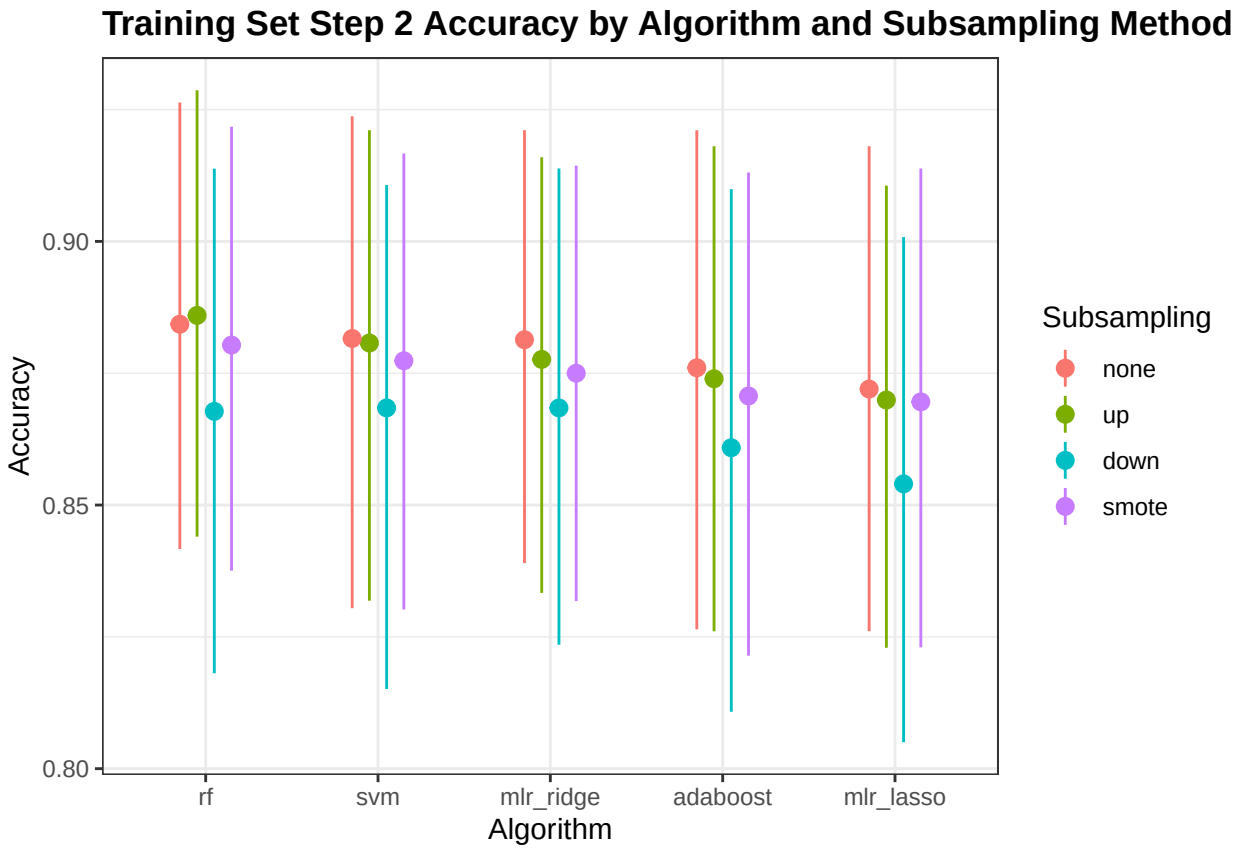


Figure 4.10: Training Set Step 2 Accuracy

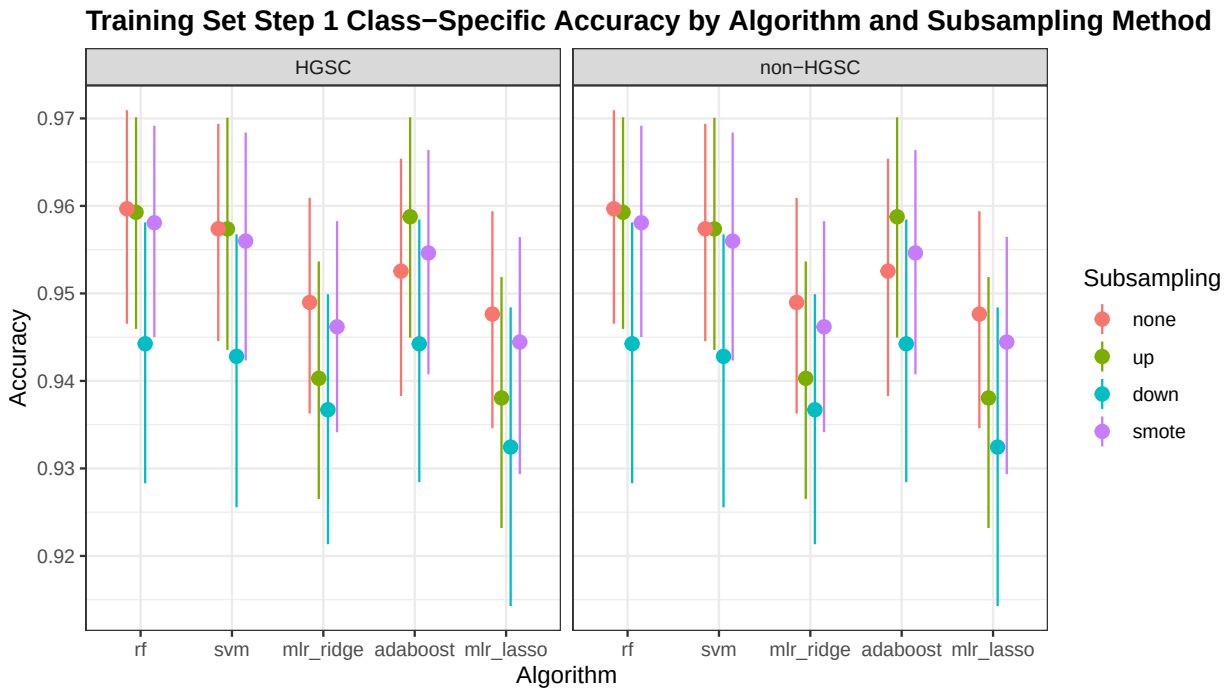


Figure 4.11: Training Set Step 1 Class-Specific Accuracy

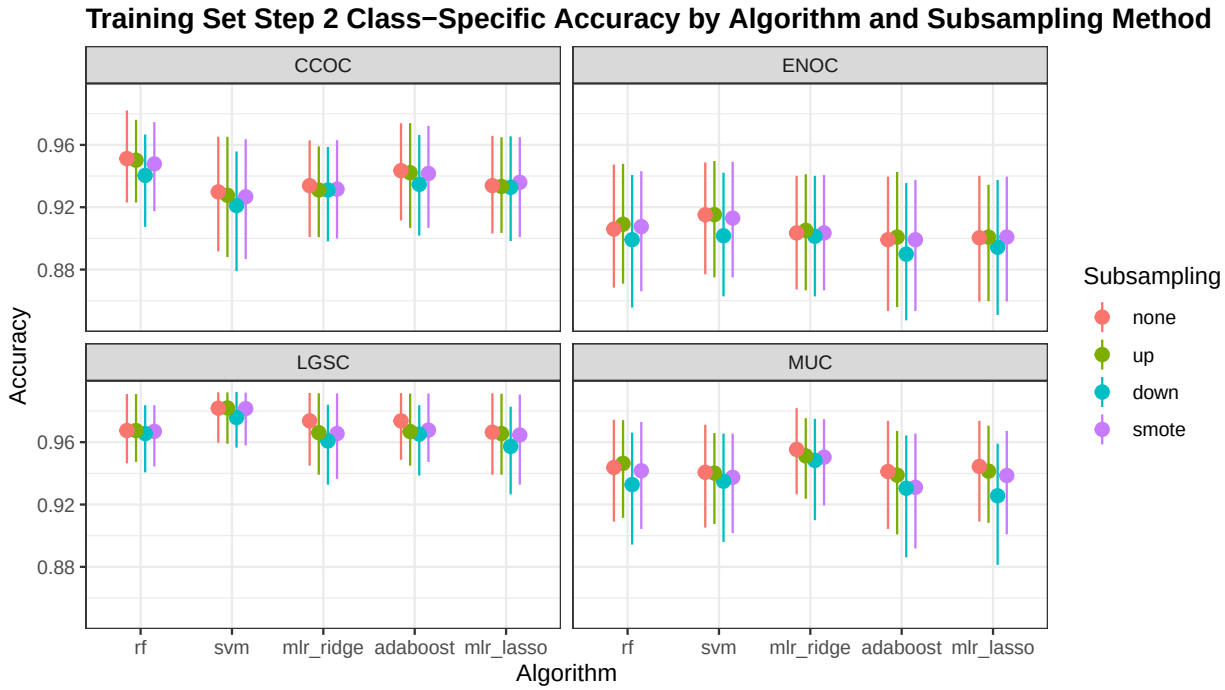


Figure 4.12: Training Set Step 2 Class-Specific Accuracy

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 | **0.96** | 0.957 | 0.949 | 0.953 | 0.948 |
| none | non-HGSC | **0.96** | 0.957 | 0.949 | 0.953 | 0.948 |
| up | HGSC | 0.959 | 0.957 | 0.94 | 0.959 | 0.938 |
| up | non-HGSC | 0.959 | 0.957 | 0.94 | 0.959 | 0.938 |
| down | HGSC | 0.944 | 0.943 | 0.937 | 0.944 | 0.932 |
| down | non-HGSC | 0.944 | 0.943 | 0.937 | 0.944 | 0.932 |
| smote | HGSC | 0.958 | 0.956 | 0.946 | 0.955 | 0.944 |
| smote | non-HGSC | 0.958 | 0.956 | 0.946 | 0.955 | 0.944 |

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 | 0.951 | 0.93 | 0.934 | 0.943 | 0.934 |
| none | ENOC | 0.906 | 0.915 | 0.904 | 0.899 | 0.9 |
| none | LGSC | 0.967 | **0.982** | 0.974 | 0.974 | 0.966 |
| none | MUC | 0.944 | 0.941 | 0.955 | 0.941 | 0.944 |
| up | CCOC | 0.95 | 0.928 | 0.931 | 0.942 | 0.933 |
| up | ENOC | 0.909 | 0.915 | 0.905 | 0.901 | 0.901 |
| up | LGSC | 0.967 | **0.982** | 0.966 | 0.967 | 0.966 |
| up | MUC | 0.946 | 0.94 | 0.951 | 0.939 | 0.941 |
| down | CCOC | 0.94 | 0.921 | 0.931 | 0.935 | 0.933 |
| down | ENOC | 0.899 | 0.902 | 0.901 | 0.89 | 0.894 |
| down | LGSC | 0.966 | 0.976 | 0.961 | 0.965 | 0.957 |
| down | MUC | 0.933 | 0.935 | 0.948 | 0.93 | 0.926 |
| smote | CCOC | 0.948 | 0.927 | 0.932 | 0.942 | 0.936 |
| smote | ENOC | 0.908 | 0.913 | 0.904 | 0.899 | 0.901 |
| smote | LGSC | 0.967 | **0.982** | 0.966 | 0.968 | 0.965 |
| smote | MUC | 0.942 | 0.938 | 0.95 | 0.931 | 0.939 |

### 4.2.2 F1-Score

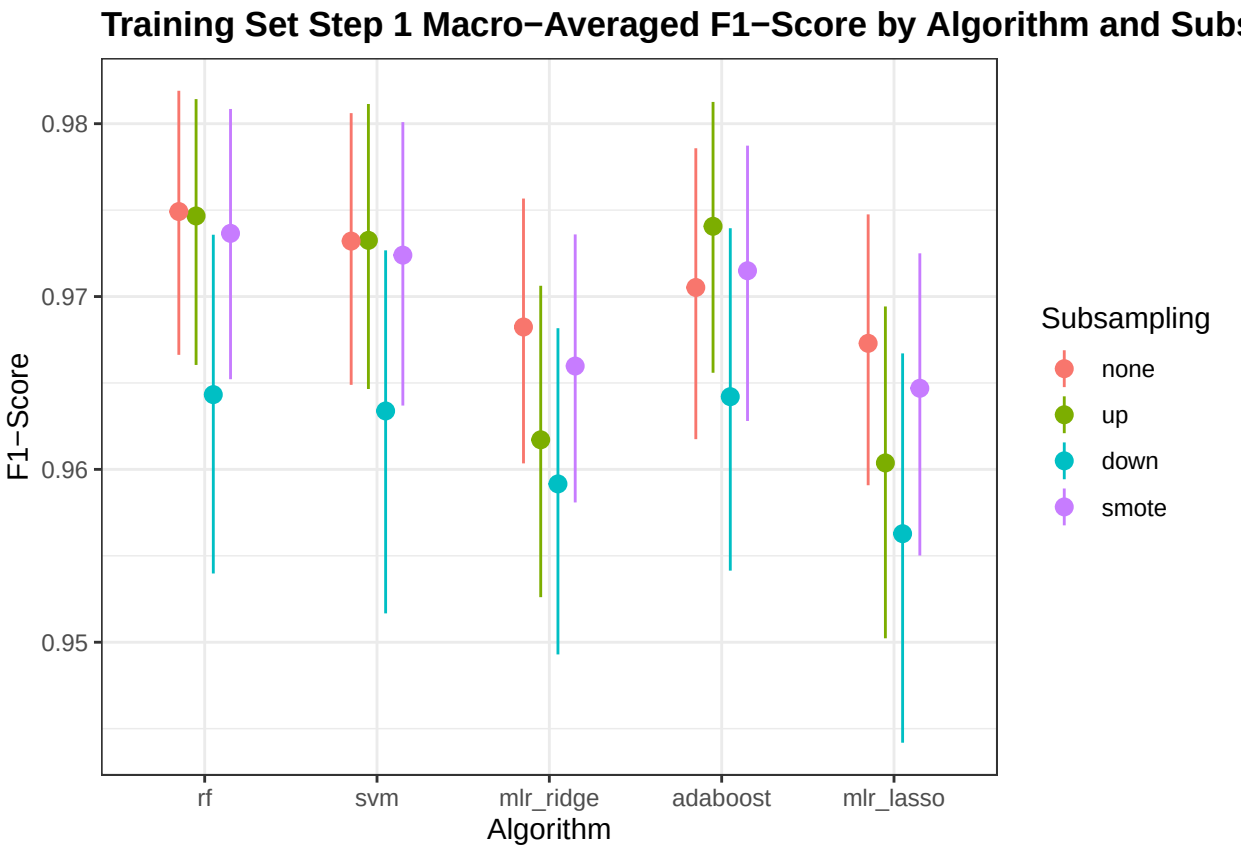


Figure 4.13: Training Set Step 1 F1-Score

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 | **0.975** | 0.973 | 0.968 | 0.971 | 0.967 |
| up | **0.975** | 0.973 | 0.962 | 0.974 | 0.96 |
| down | 0.964 | 0.963 | 0.959 | 0.964 | 0.956 |
| smote | 0.974 | 0.972 | 0.966 | 0.971 | 0.965 |

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 | 0.882 | **0.885** | 0.88 | 0.877 | 0.87 |
| up | 0.883 | 0.884 | 0.876 | 0.873 | 0.868 |
| down | 0.866 | 0.873 | 0.866 | 0.859 | 0.85 |
| smote | 0.879 | 0.881 | 0.874 | 0.869 | 0.866 |

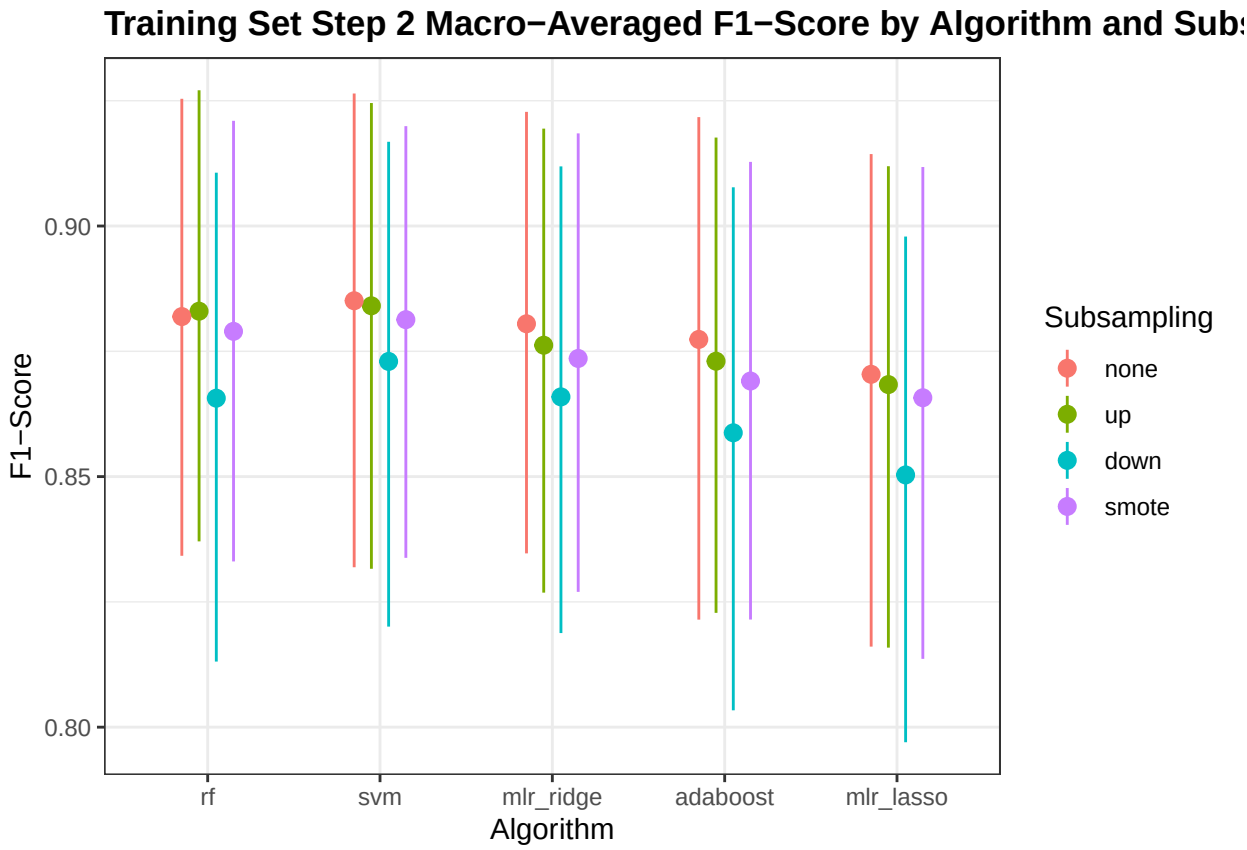


Figure 4.14: Training Set Step 2 F1-Score

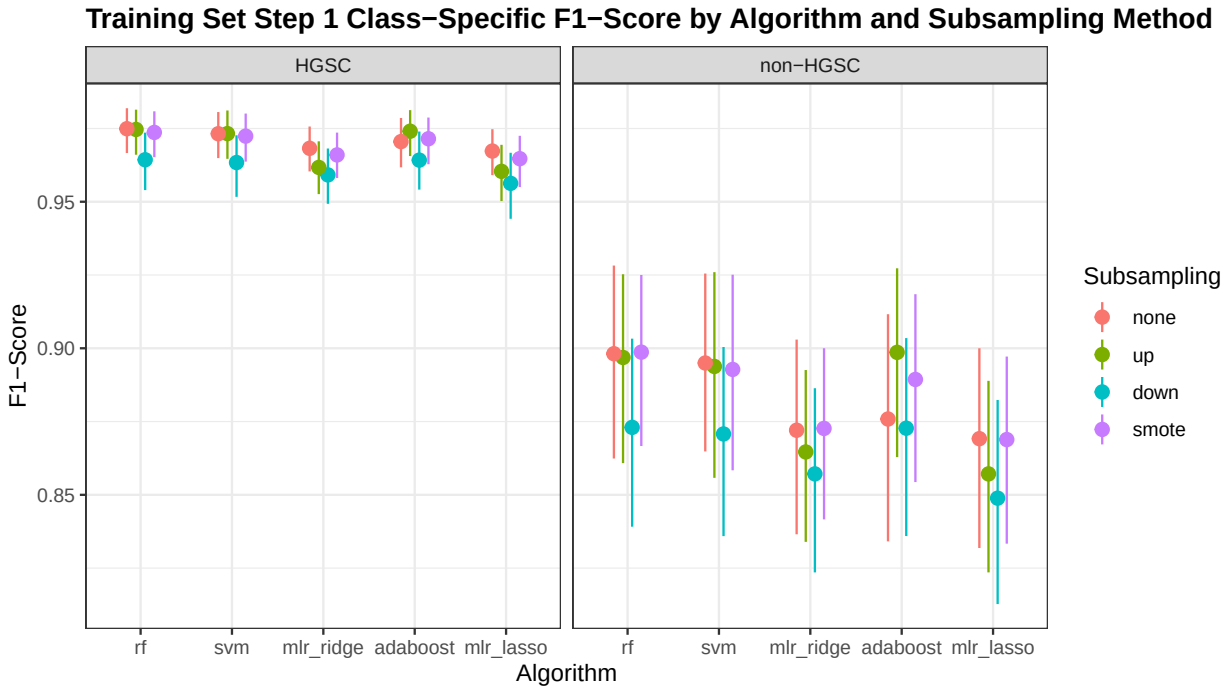


Figure 4.15: Training Set Step 1 Class-Specific F1-Score

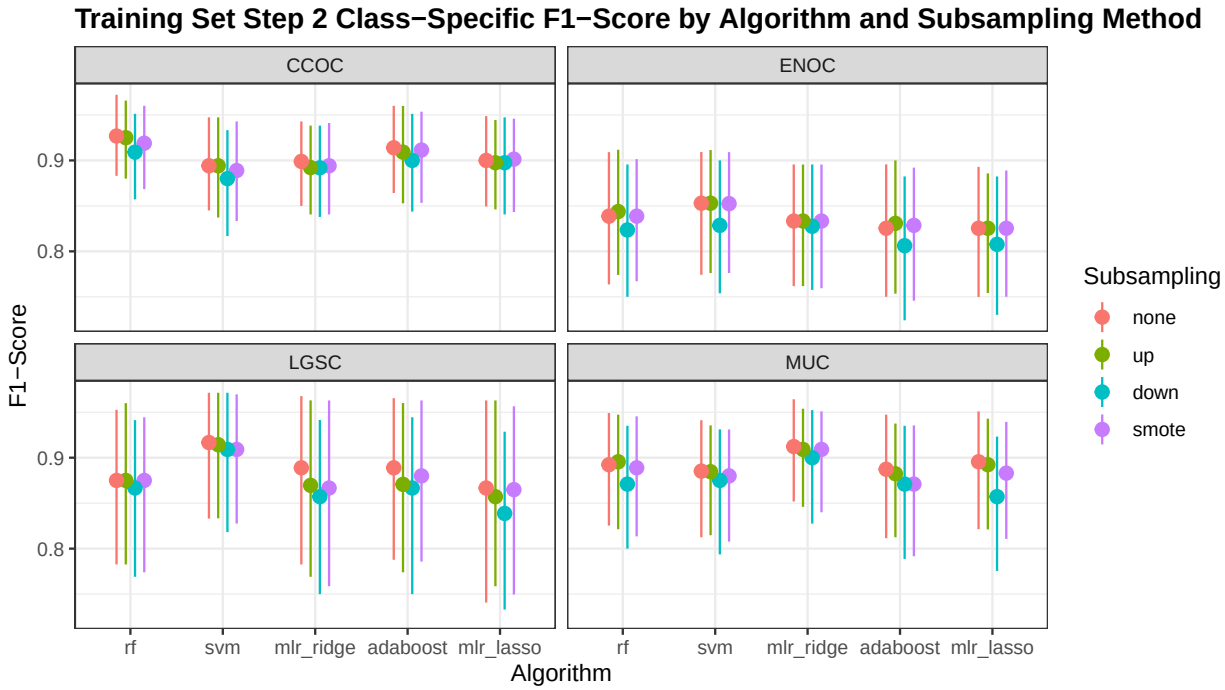


Figure 4.16: Training Set Step 2 Class-Specific F1-Score

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{0.975} | 0.973 | 0.968 | 0.971 | 0.967 |
| none | non-HGSC | 0.898 | 0.895 | 0.872 | 0.876 | 0.869 |
| up | HGSC | \textbf{0.975} | 0.973 | 0.962 | 0.974 | 0.96 |
| up | non-HGSC | 0.897 | 0.894 | 0.865 | 0.899 | 0.857 |
| down | HGSC | 0.964 | 0.963 | 0.959 | 0.964 | 0.956 |
| down | non-HGSC | 0.873 | 0.871 | 0.857 | 0.873 | 0.849 |
| smote | HGSC | 0.974 | 0.972 | 0.966 | 0.971 | 0.965 |
| smote | non-HGSC | 0.899 | 0.893 | 0.873 | 0.889 | 0.869 |

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{0.927} | 0.894 | 0.899 | 0.914 | 0.9 |
| none | ENOC | 0.839 | 0.853 | 0.833 | 0.825 | 0.825 |
| none | LGSC | 0.875 | 0.917 | 0.889 | 0.889 | 0.867 |
| none | MUC | 0.892 | 0.885 | 0.912 | 0.887 | 0.896 |
| up | CCOC | 0.925 | 0.894 | 0.892 | 0.909 | 0.897 |
| up | ENOC | 0.844 | 0.853 | 0.833 | 0.831 | 0.825 |
| up | LGSC | 0.875 | 0.914 | 0.87 | 0.871 | 0.857 |
| up | MUC | 0.896 | 0.885 | 0.909 | 0.882 | 0.892 |
| down | CCOC | 0.909 | 0.88 | 0.892 | 0.9 | 0.897 |
| down | ENOC | 0.824 | 0.829 | 0.828 | 0.806 | 0.808 |
| down | LGSC | 0.867 | 0.909 | 0.857 | 0.867 | 0.839 |
| down | MUC | 0.871 | 0.875 | 0.9 | 0.871 | 0.857 |
| smote | CCOC | 0.919 | 0.889 | 0.894 | 0.911 | 0.901 |
| smote | ENOC | 0.839 | 0.852 | 0.833 | 0.829 | 0.825 |
| smote | LGSC | 0.875 | 0.909 | 0.867 | 0.88 | 0.865 |
| smote | MUC | 0.889 | 0.88 | 0.909 | 0.871 | 0.883 |

Showing 1 to 16 of 16 entries

Previous 1 Next

### 4.2.3 Kappa

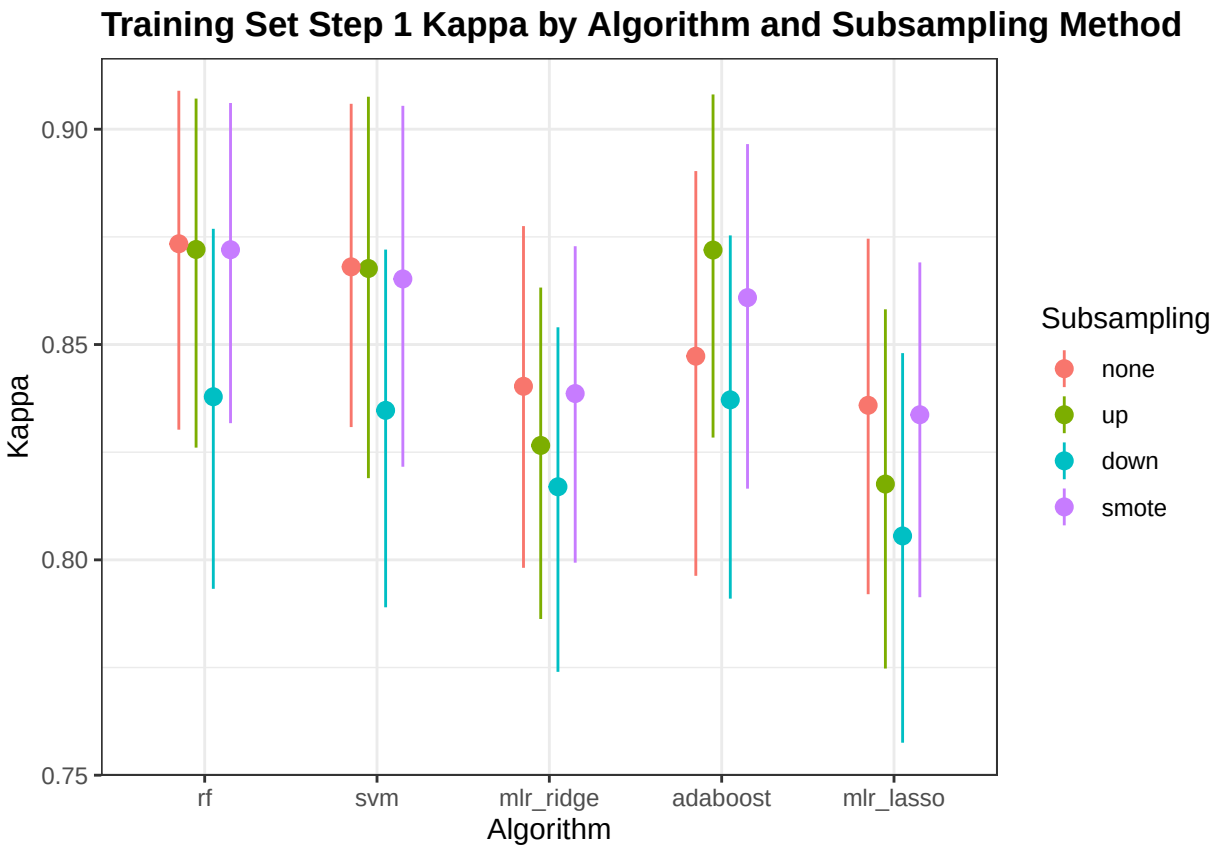


Figure 4.17: Training Set Step 1 Kappa

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

| sampling | rf | svm | mlr_ridge | adaboost | mlr_lasso |
|---|---|---|---|---|---|
| none | **0.873** | 0.868 | 0.84 | 0.847 | 0.836 |
| up | 0.872 | 0.868 | 0.827 | 0.872 | 0.818 |
| down | 0.838 | 0.835 | 0.817 | 0.837 | 0.806 |
| smote | 0.872 | 0.865 | 0.839 | 0.861 | 0.834 |

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

| sampling | rf | svm | mlr_ridge | adaboost | mlr_lasso |
|---|---|---|---|---|---|
| none | 0.841 | 0.836 | 0.836 | 0.829 | 0.824 |
| up | **0.842** | 0.836 | 0.832 | 0.827 | 0.821 |
| down | 0.818 | 0.819 | 0.82 | 0.808 | 0.801 |
| smote | 0.835 | 0.831 | 0.827 | 0.822 | 0.82 |

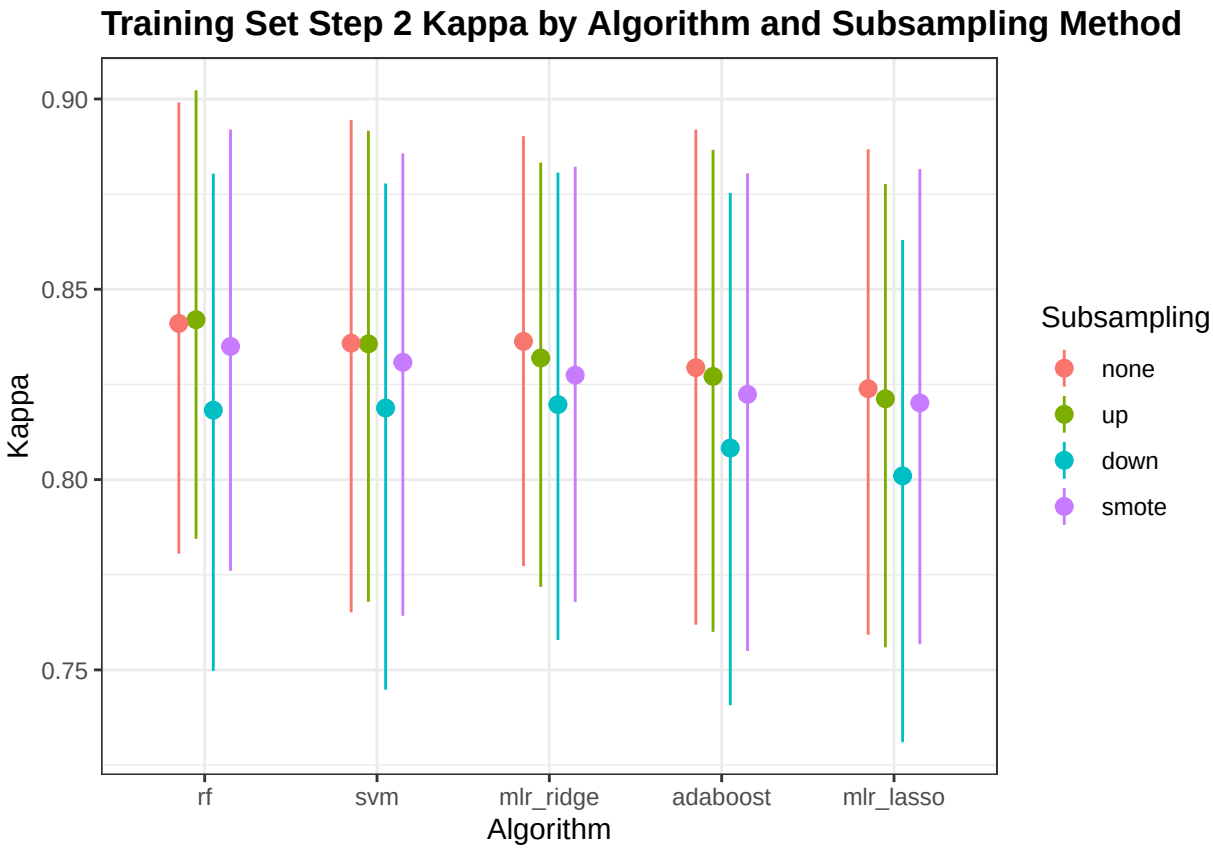


Figure 4.18: Training Set Step 2 Kappa

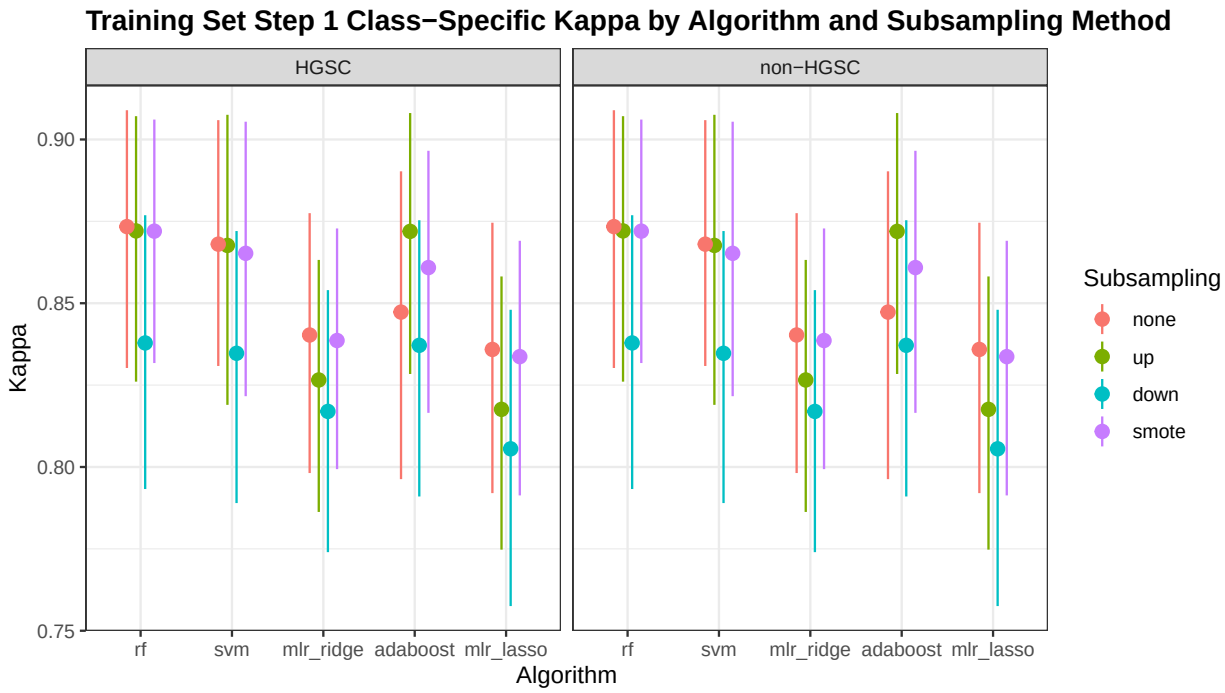


Figure 4.19: Training Set Step 1 Class-Specific Kappa

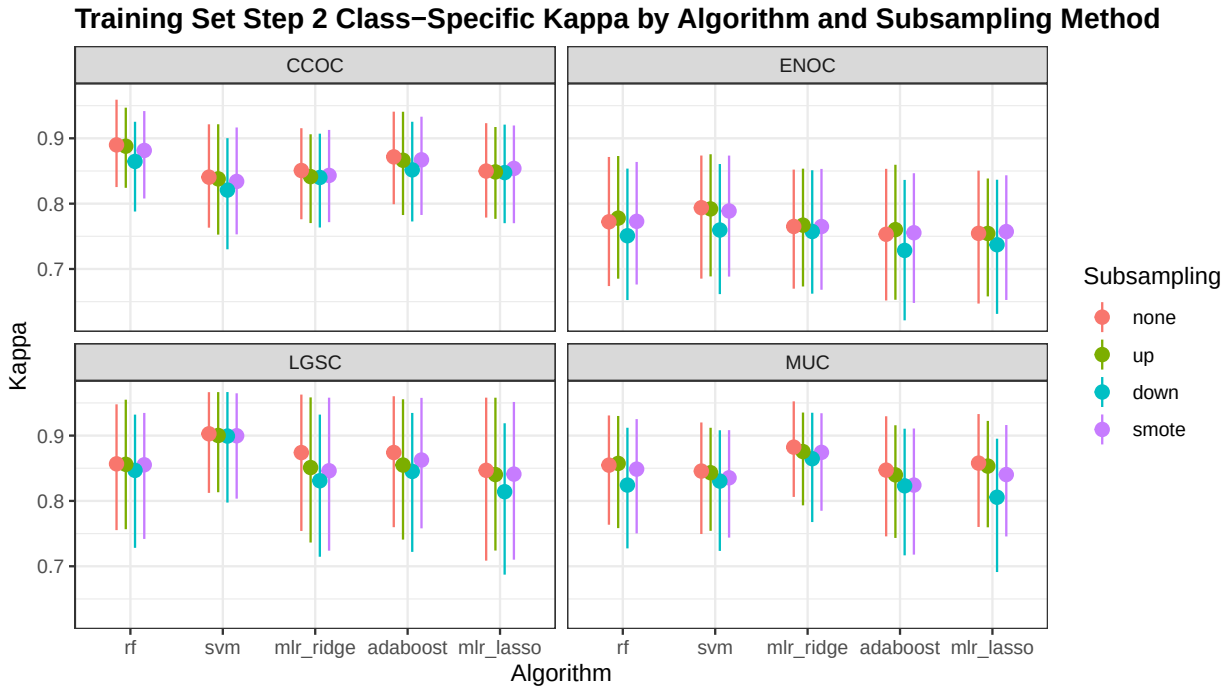


Figure 4.20: Training Set Step 2 Class-Specific Kappa

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 | **0.873** | 0.868 | 0.84 | 0.847 | 0.836 |
| none | non-HGSC | **0.873** | 0.868 | 0.84 | 0.847 | 0.836 |
| up | HGSC | 0.872 | 0.868 | 0.827 | 0.872 | 0.818 |
| up | non-HGSC | 0.872 | 0.868 | 0.827 | 0.872 | 0.818 |
| down | HGSC | 0.838 | 0.835 | 0.817 | 0.837 | 0.806 |
| down | non-HGSC | 0.838 | 0.835 | 0.817 | 0.837 | 0.806 |
| smote | HGSC | 0.872 | 0.865 | 0.839 | 0.861 | 0.834 |
| smote | non-HGSC | 0.872 | 0.865 | 0.839 | 0.861 | 0.834 |

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 | 0.89 | 0.841 | 0.851 | 0.872 | 0.85 |
| none | ENOC | 0.772 | 0.794 | 0.765 | 0.753 | 0.755 |
| none | LGSC | 0.857 | **0.903** | 0.874 | 0.874 | 0.847 |
| none | MUC | 0.855 | 0.846 | 0.882 | 0.847 | 0.858 |
| up | CCOC | 0.888 | 0.838 | 0.841 | 0.866 | 0.849 |
| up | ENOC | 0.778 | 0.792 | 0.767 | 0.76 | 0.754 |
| up | LGSC | 0.856 | 0.9 | 0.851 | 0.855 | 0.84 |
| up | MUC | 0.857 | 0.843 | 0.876 | 0.84 | 0.853 |
| down | CCOC | 0.865 | 0.821 | 0.84 | 0.852 | 0.848 |
| down | ENOC | 0.751 | 0.76 | 0.757 | 0.728 | 0.737 |
| down | LGSC | 0.847 | 0.899 | 0.831 | 0.845 | 0.814 |
| down | MUC | 0.824 | 0.83 | 0.865 | 0.823 | 0.806 |
| smote | CCOC | 0.881 | 0.834 | 0.843 | 0.867 | 0.854 |
| smote | ENOC | 0.773 | 0.789 | 0.765 | 0.755 | 0.757 |
| smote | LGSC | 0.855 | 0.9 | 0.846 | 0.863 | 0.841 |
| smote | MUC | 0.849 | 0.835 | 0.874 | 0.824 | 0.84 |

### 4.2.4 G-mean

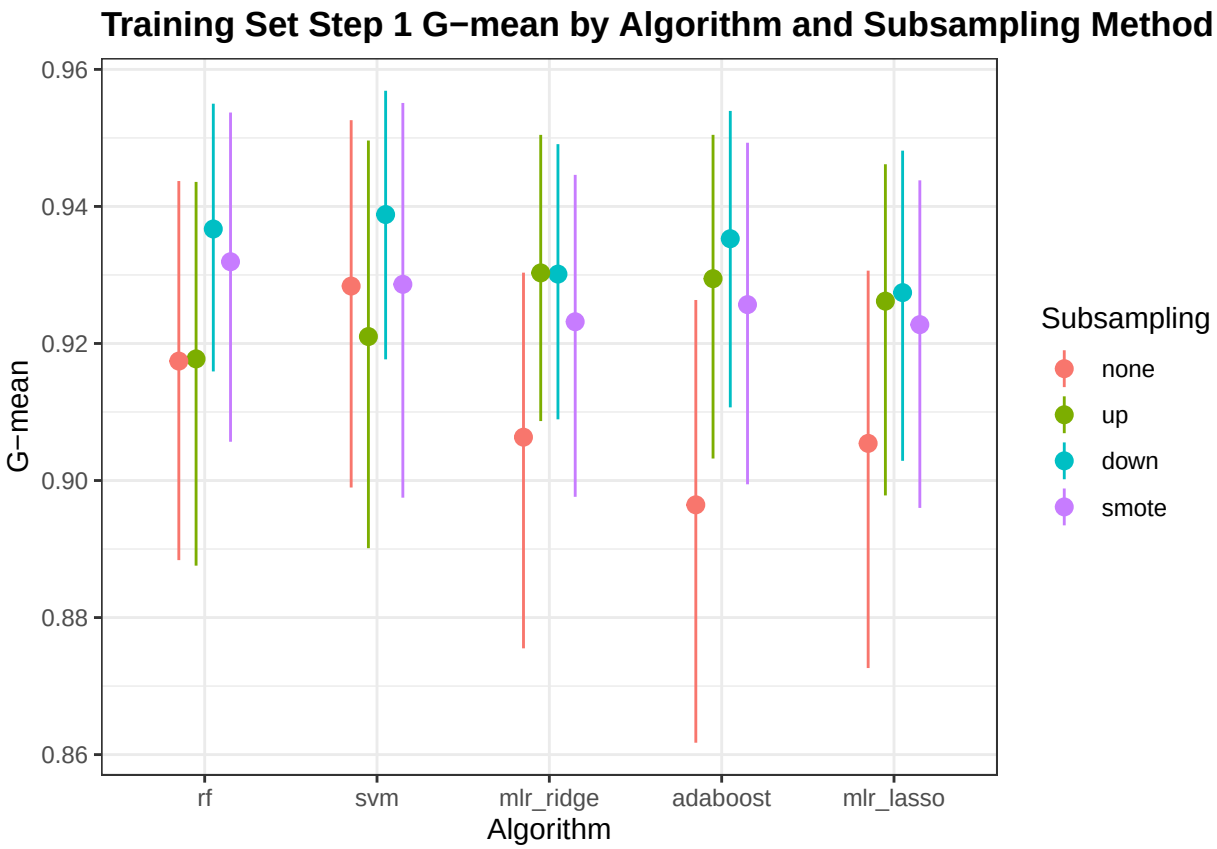


Figure 4.21: Training Set Step 1 G-mean

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

| sampling | rf | svm | mlr_ridge | adaboost | mlr_lasso |
|---|---|---|---|---|---|
| none | 0.917 | 0.928 | 0.906 | 0.896 | 0.905 |
| up | 0.918 | 0.921 | 0.93 | 0.929 | 0.926 |
| down | 0.937 | **0.939** | 0.93 | 0.935 | 0.927 |
| smote | 0.932 | 0.929 | 0.923 | 0.926 | 0.923 |

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

| sampling | rf | svm | mlr_ridge | adaboost | mlr_lasso |
|---|---|---|---|---|---|
| none | 0.885 | 0.882 | 0.883 | 0.877 | 0.869 |
| up | **0.887** | 0.881 | 0.882 | 0.878 | 0.874 |
| down | 0.874 | 0.874 | 0.877 | 0.865 | 0.859 |
| smote | 0.885 | 0.88 | 0.881 | 0.875 | 0.871 |

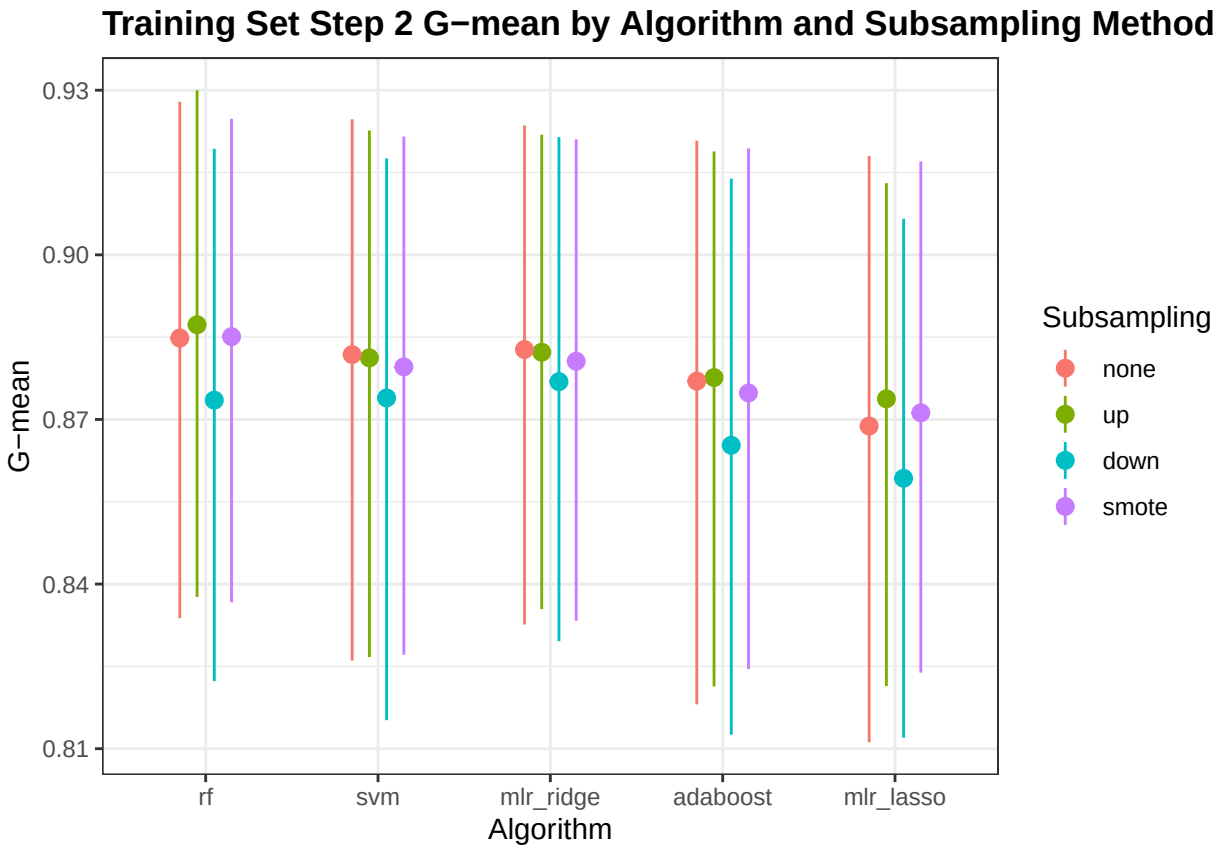


Figure 4.22: Training Set Step 2 G-mean

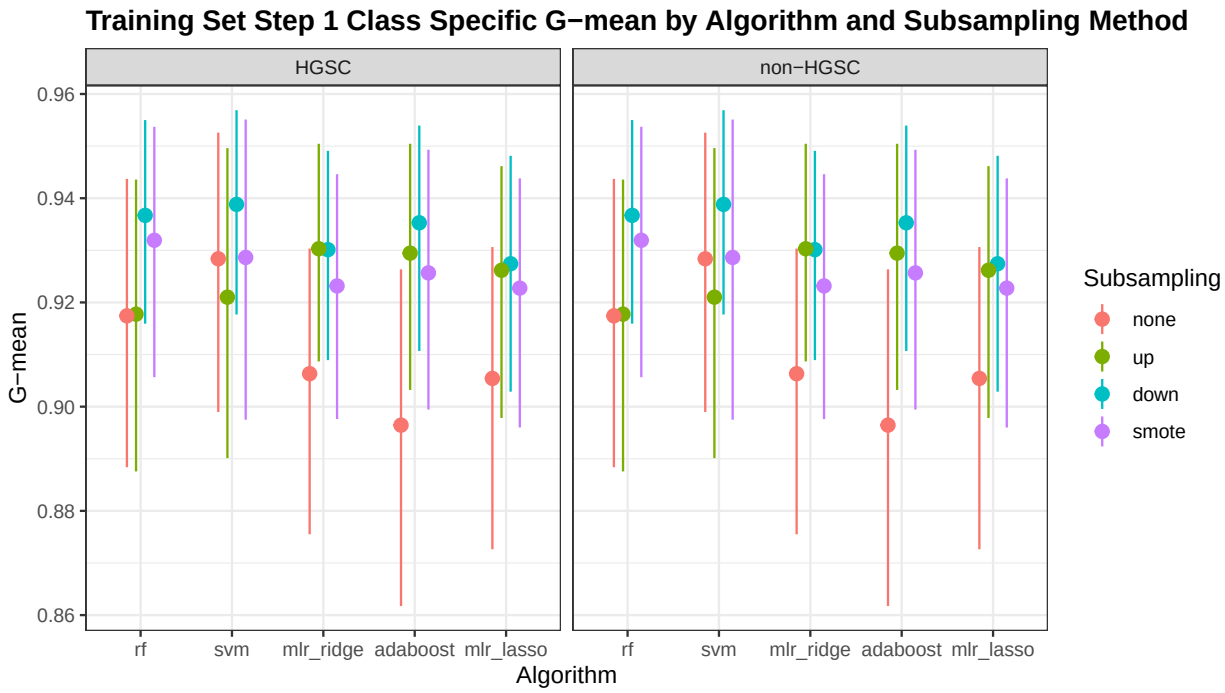


Figure 4.23: Training Set Step 1 Class-Specific G-mean

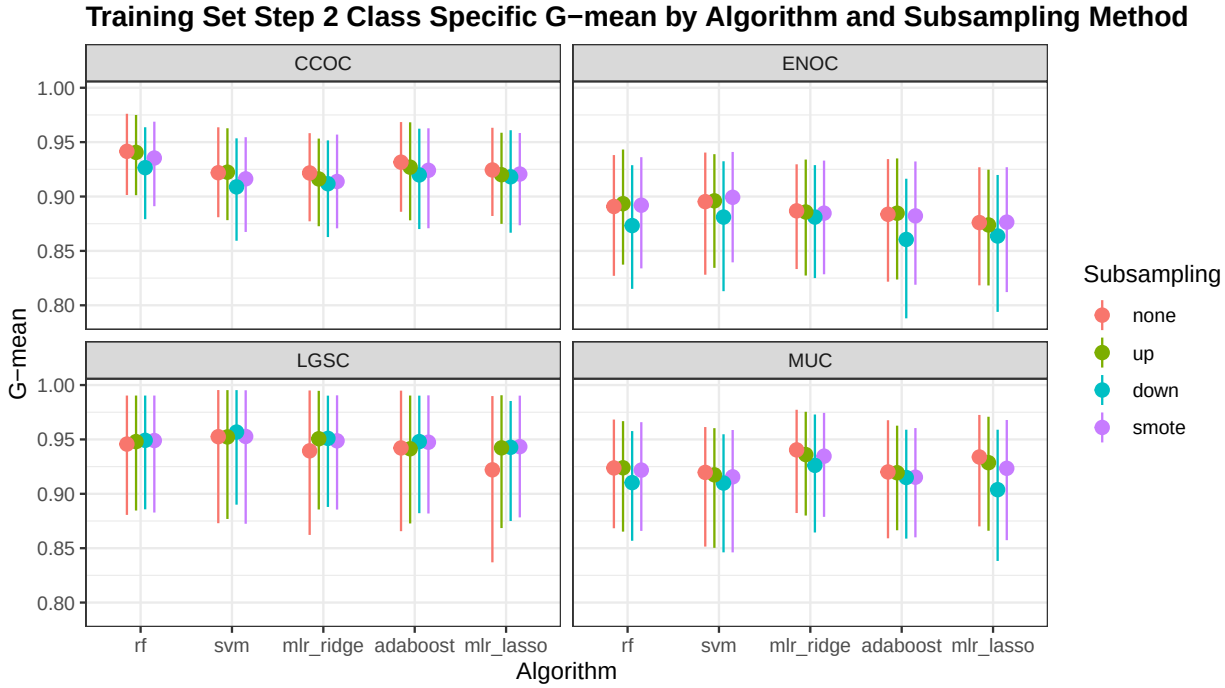


Figure 4.24: Training Set Step 2 Class-Specific G-mean

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 | 0.917 | 0.928 | 0.906 | 0.896 | 0.905 |
| none | non-HGSC | 0.917 | 0.928 | 0.906 | 0.896 | 0.905 |
| up | HGSC | 0.918 | 0.921 | 0.93 | 0.929 | 0.926 |
| up | non-HGSC | 0.918 | 0.921 | 0.93 | 0.929 | 0.926 |
| down | HGSC | 0.937 | **0.939** | 0.93 | 0.935 | 0.927 |
| down | non-HGSC | 0.937 | **0.939** | 0.93 | 0.935 | 0.927 |
| smote | HGSC | 0.932 | 0.929 | 0.923 | 0.926 | 0.923 |
| smote | non-HGSC | 0.932 | 0.929 | 0.923 | 0.926 | 0.923 |

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 | 0.942 | 0.922 | 0.922 | 0.932 | 0.924 |
| none | ENOC | 0.891 | 0.895 | 0.887 | 0.884 | 0.876 |
| none | LGSC | 0.946 | 0.953 | 0.94 | 0.942 | 0.922 |
| none | MUC | 0.924 | 0.92 | 0.94 | 0.92 | 0.934 |
| up | CCOC | 0.941 | 0.922 | 0.916 | 0.927 | 0.92 |
| up | ENOC | 0.893 | 0.896 | 0.886 | 0.885 | 0.874 |
| up | LGSC | 0.948 | 0.952 | 0.951 | 0.941 | 0.942 |
| up | MUC | 0.924 | 0.917 | 0.936 | 0.919 | 0.929 |
| down | CCOC | 0.927 | 0.909 | 0.912 | 0.92 | 0.918 |
| down | ENOC | 0.873 | 0.881 | 0.881 | 0.861 | 0.864 |
| down | LGSC | 0.949 | **0.957** | 0.951 | 0.948 | 0.943 |
| down | MUC | 0.91 | 0.91 | 0.926 | 0.915 | 0.904 |
| smote | CCOC | 0.935 | 0.916 | 0.914 | 0.924 | 0.921 |
| smote | ENOC | 0.892 | 0.899 | 0.885 | 0.882 | 0.876 |
| smote | LGSC | 0.949 | 0.953 | 0.949 | 0.947 | 0.943 |
| smote | MUC | 0.922 | 0.916 | 0.935 | 0.915 | 0.923 |

Table 4.23: CS1 Set Accuracy by Algorithm and Subsampling Method

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

## 4.3 CS1 Set

### 4.3.1 Accuracy

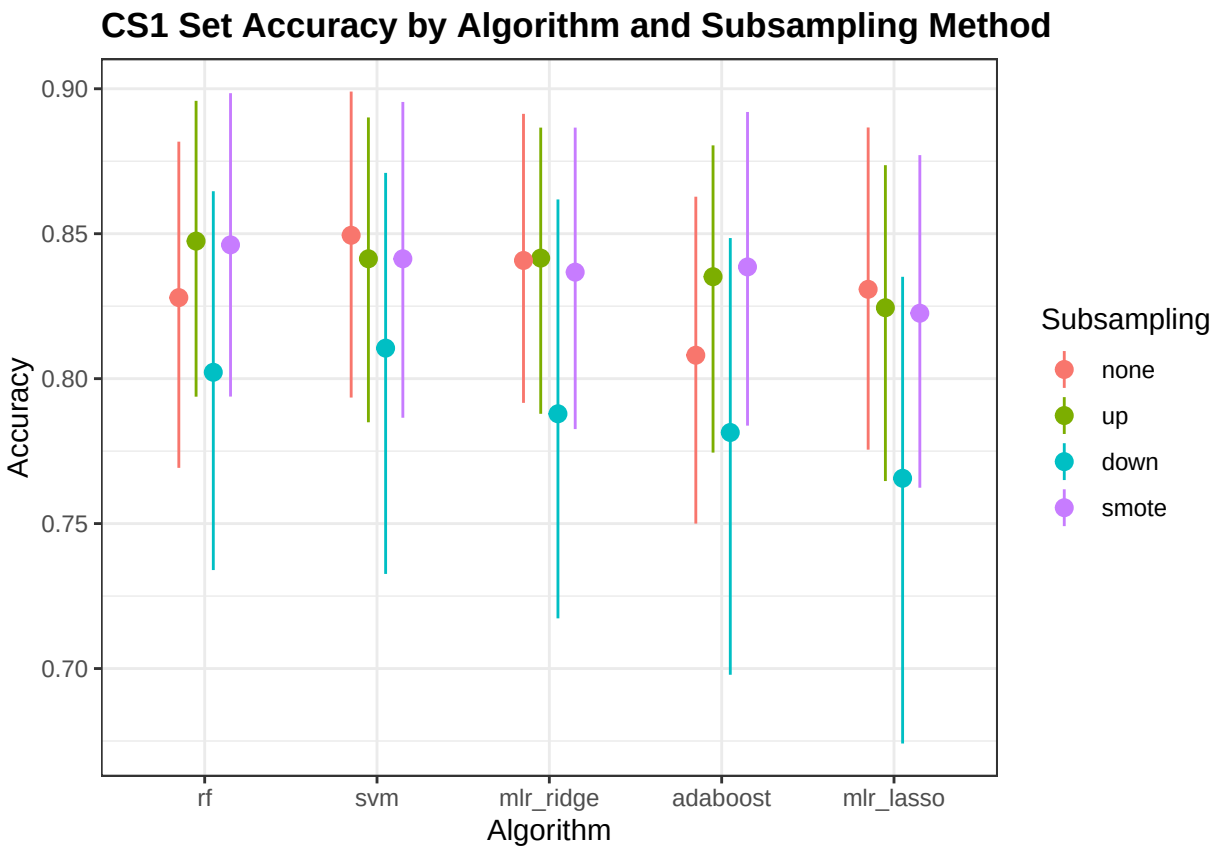


Figure 4.25: CS1 Set Accuracy

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

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

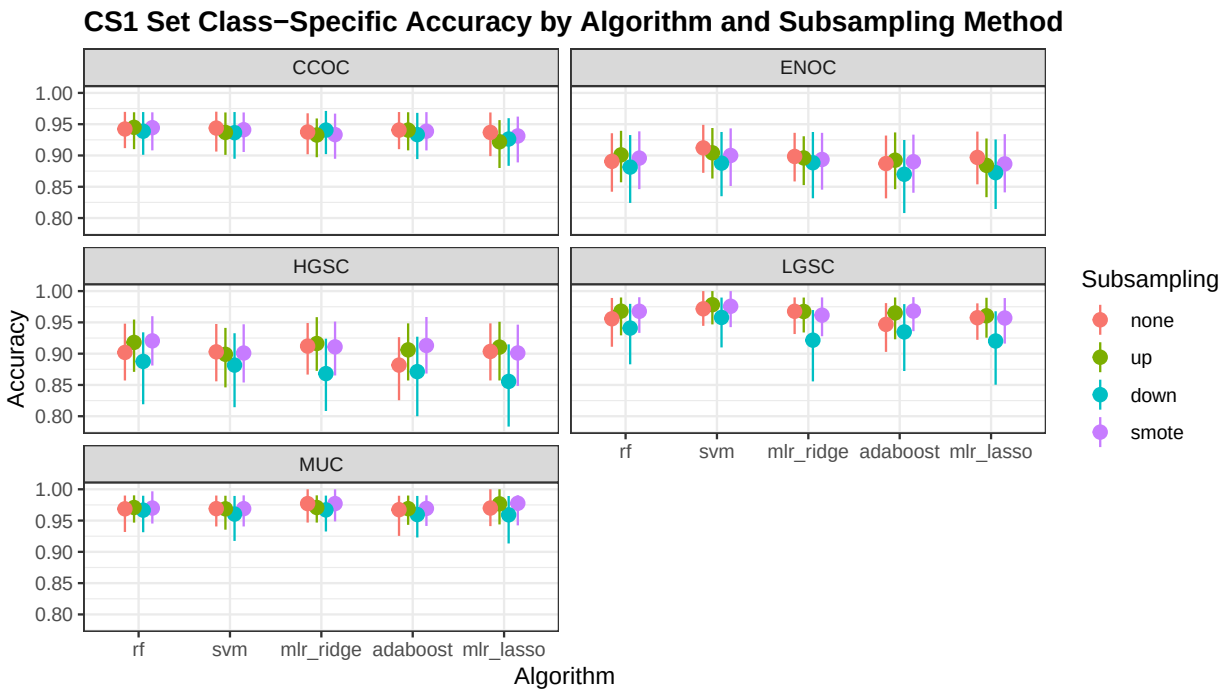


Figure 4.26: CS1 Set Class-Specific Accuracy

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

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

### 4.3.2 F1-Score

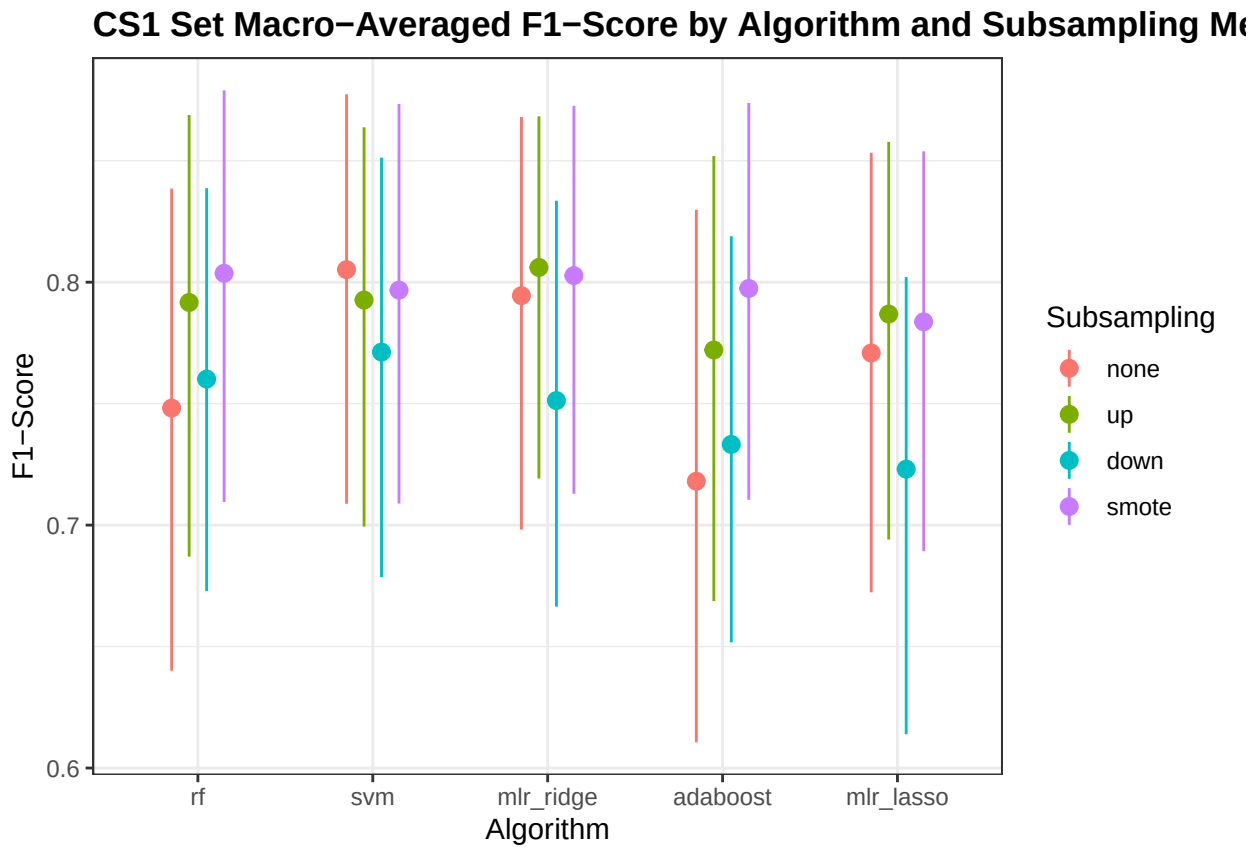


Figure 4.27: CS1 Set F1-Score

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

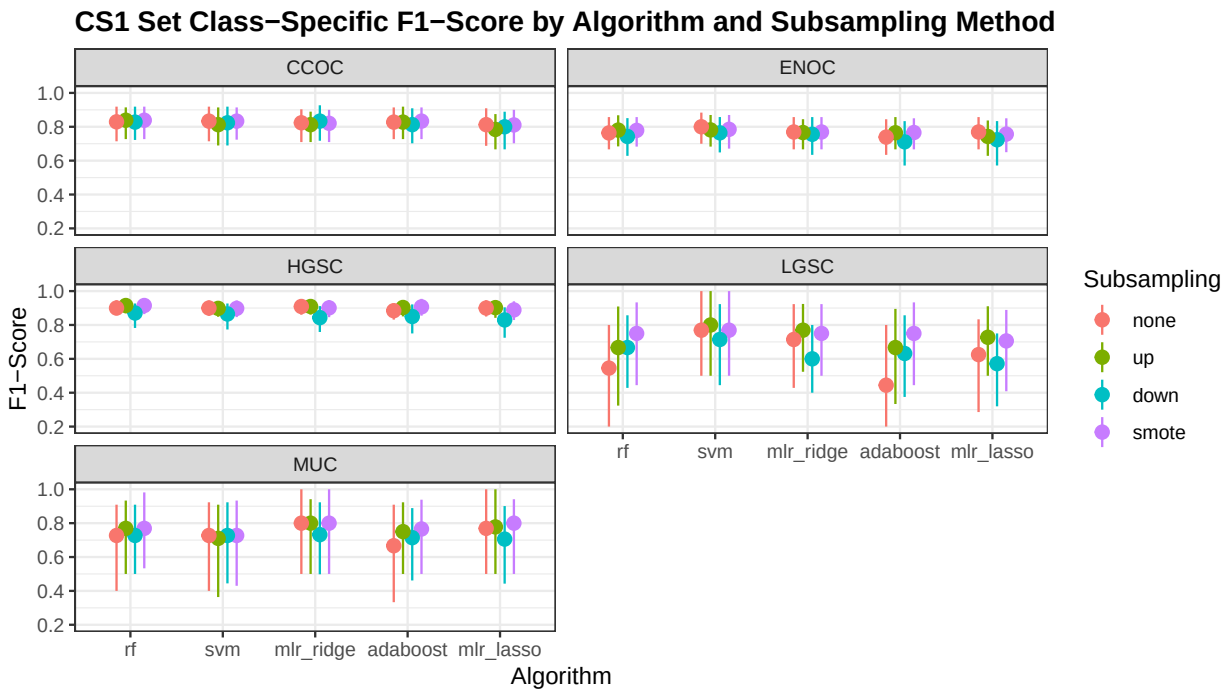


Figure 4.28: CS1 Set Class-Specific F1-Score

Table 4.27: CS1 Set Kappa by Algorithm and Subsampling Method

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

### 4.3.3 Kappa

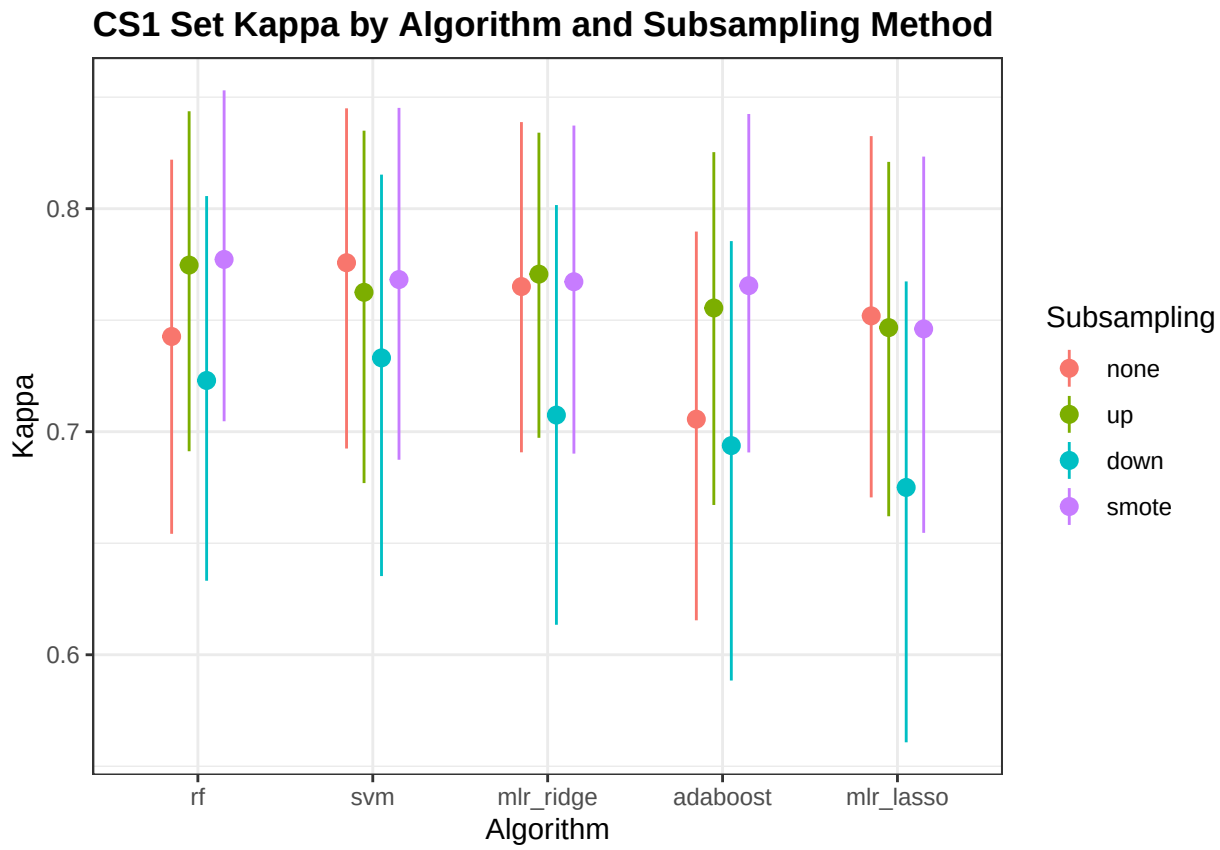


Figure 4.29: CS1 Set Kappa

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

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

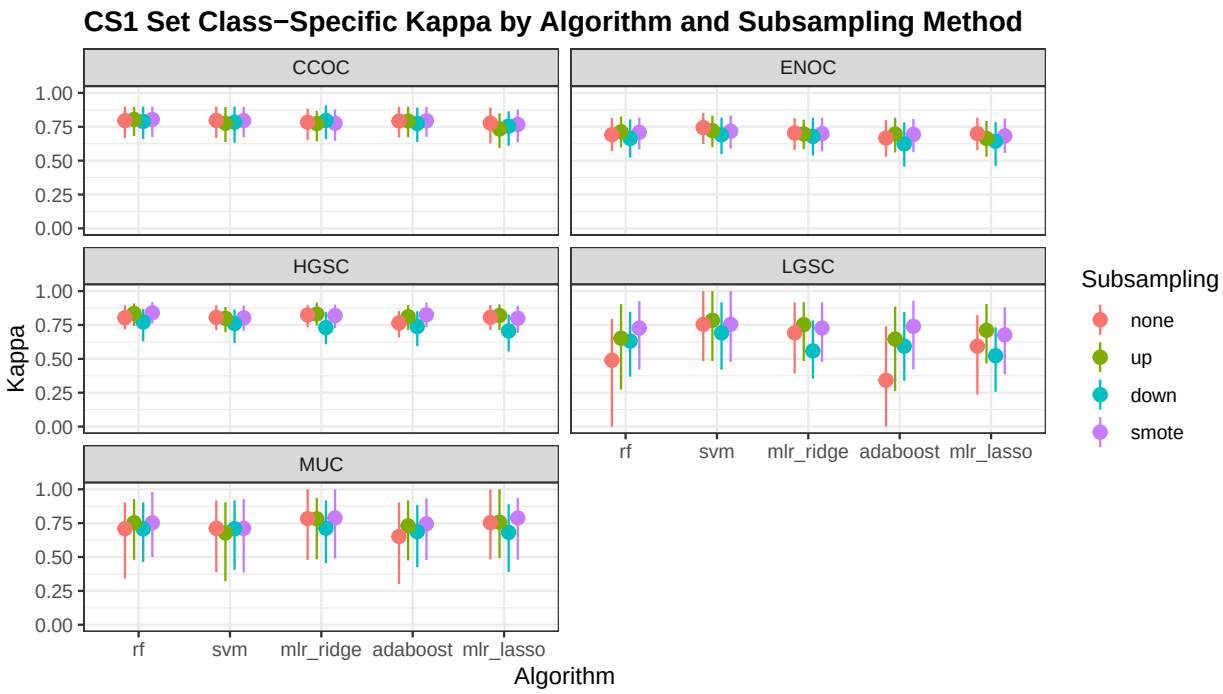


Figure 4.30: CS1 Set Class-Specific Kappa

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

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

### 4.3.4 G-mean

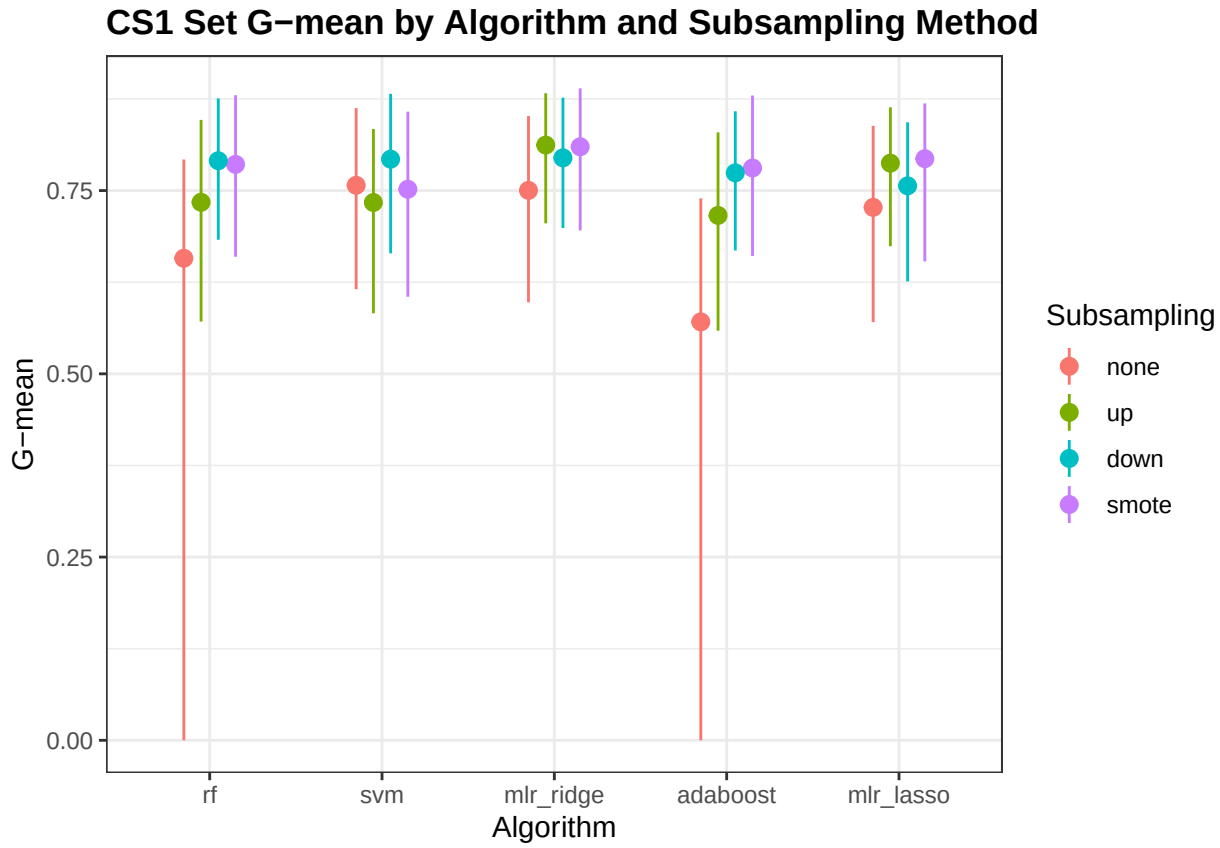


Figure 4.31: CS1 Set G-mean

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

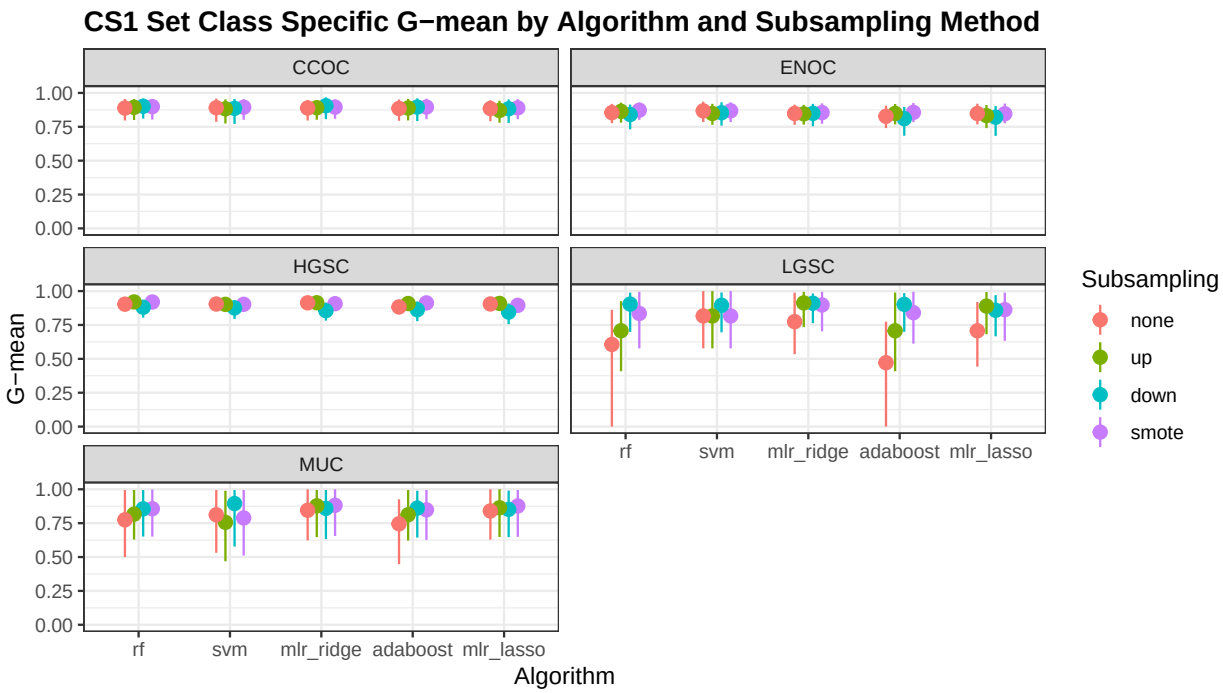


Figure 4.32: CS1 Set Class-Specific G-mean

Table 4.31: CS2 Set Accuracy by Algorithm and Subsampling Method

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

## 4.4 CS2 Set

### 4.4.1 Accuracy

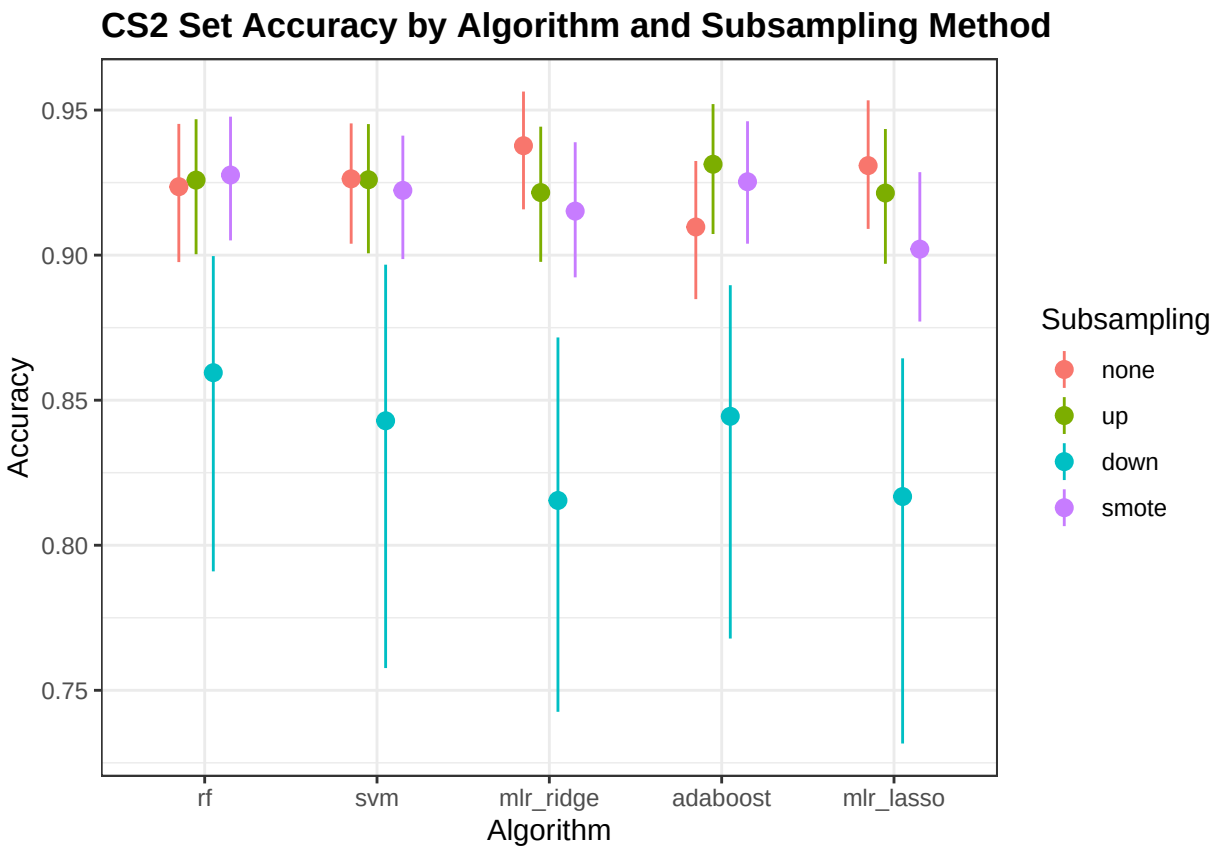


Figure 4.33: CS2 Set Accuracy

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

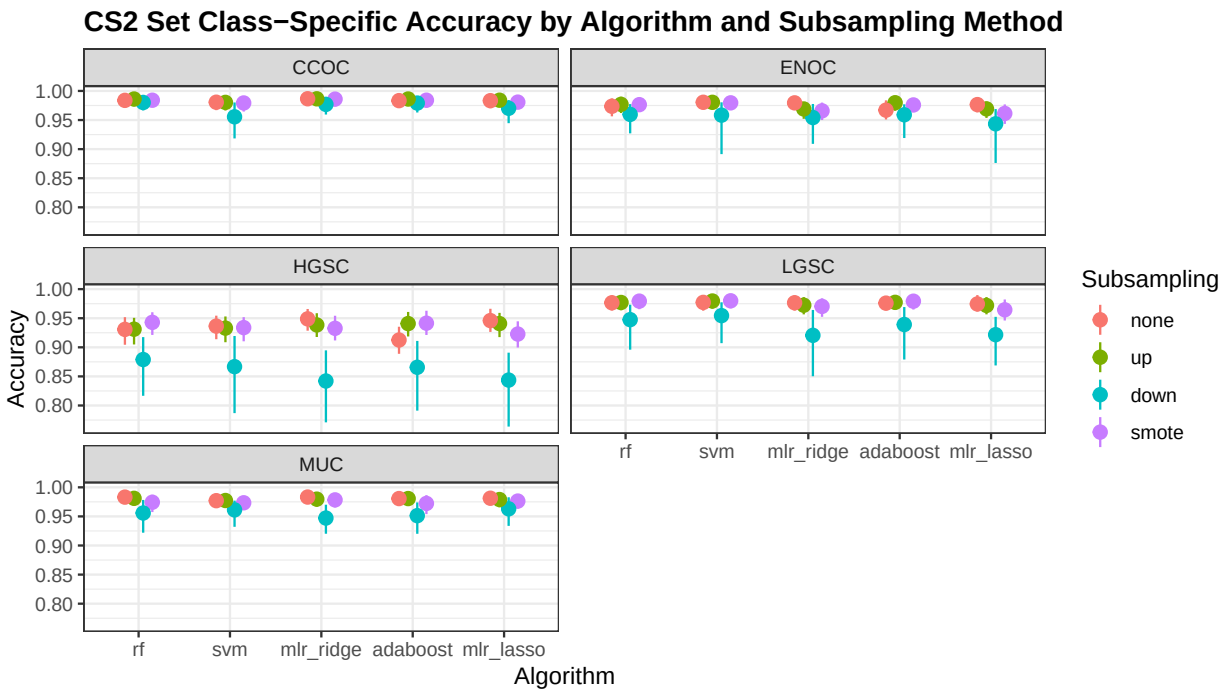


Figure 4.34: CS2 Set Class-Specific Accuracy

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

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

### 4.4.2 F1-Score

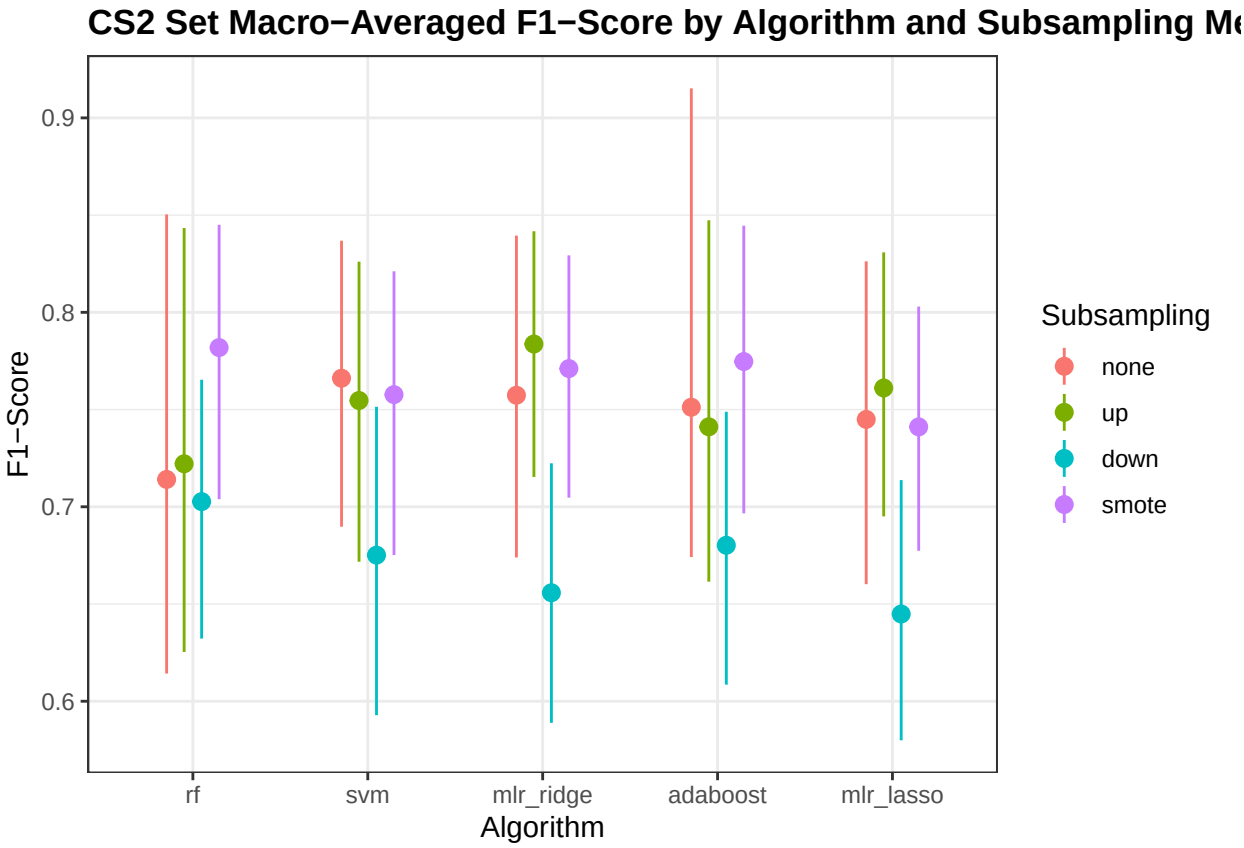


Figure 4.35: CS2 Set F1-Score

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

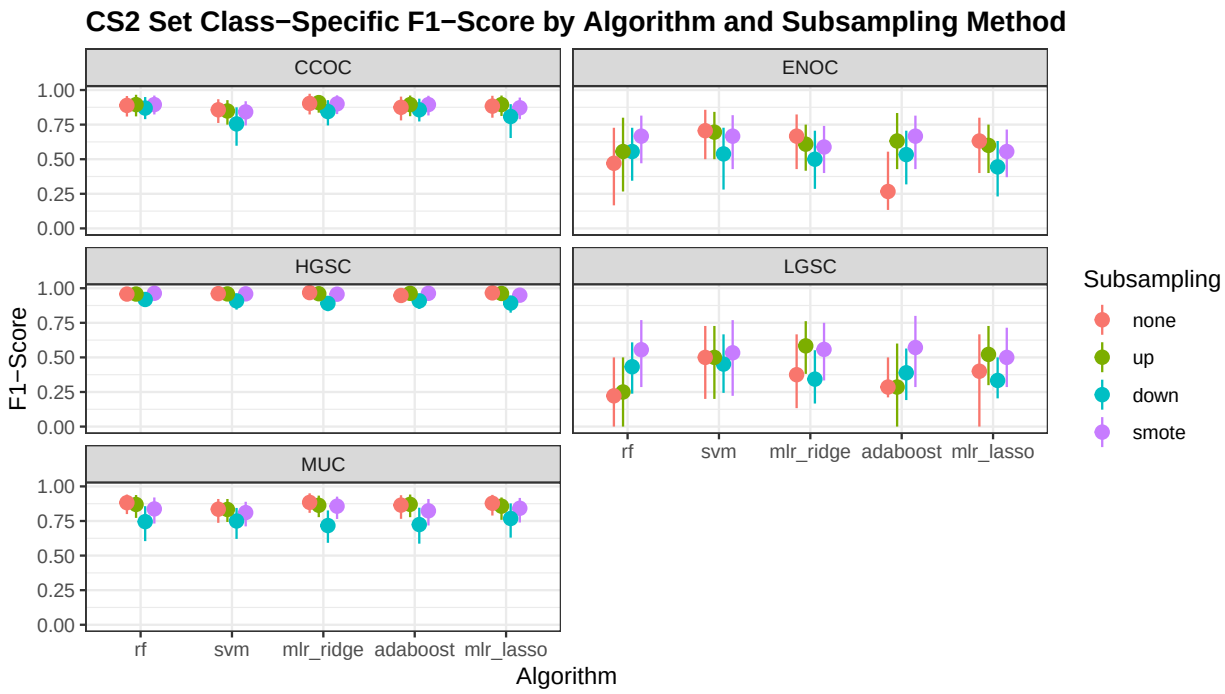


Figure 4.36: CS2 Set Class-Specific F1-Score

Table 4.35: CS2 Set Kappa by Algorithm and Subsampling Method

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

### 4.4.3 Kappa

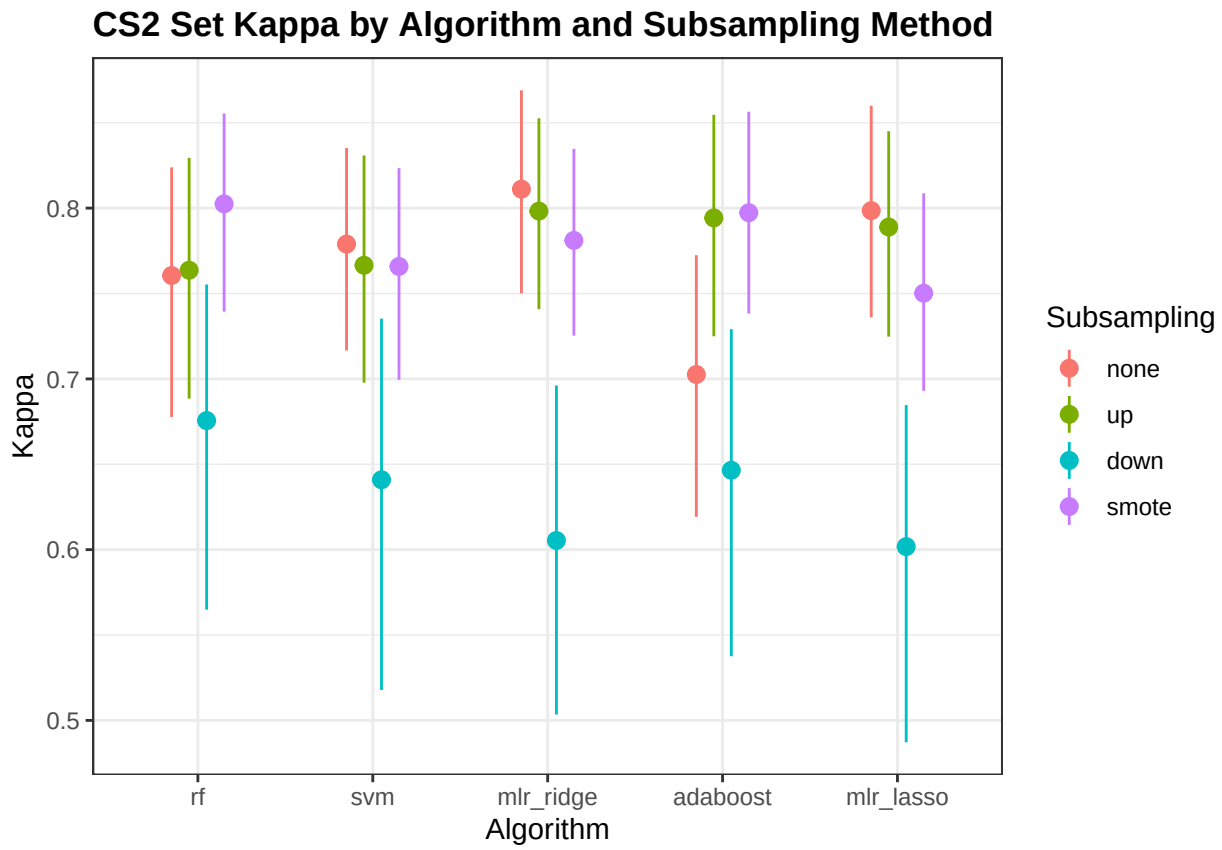


Figure 4.37: CS2 Set Kappa

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

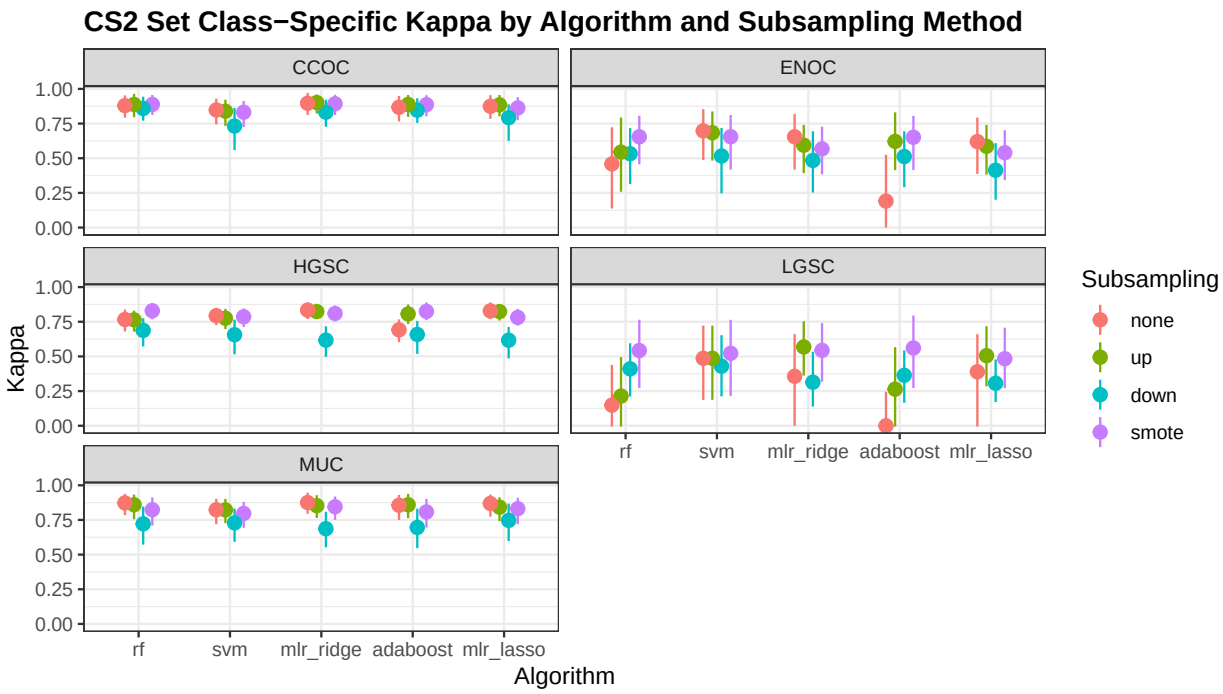


Figure 4.38: CS2 Set Class-Specific Kappa

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

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

### 4.4.4 G-mean

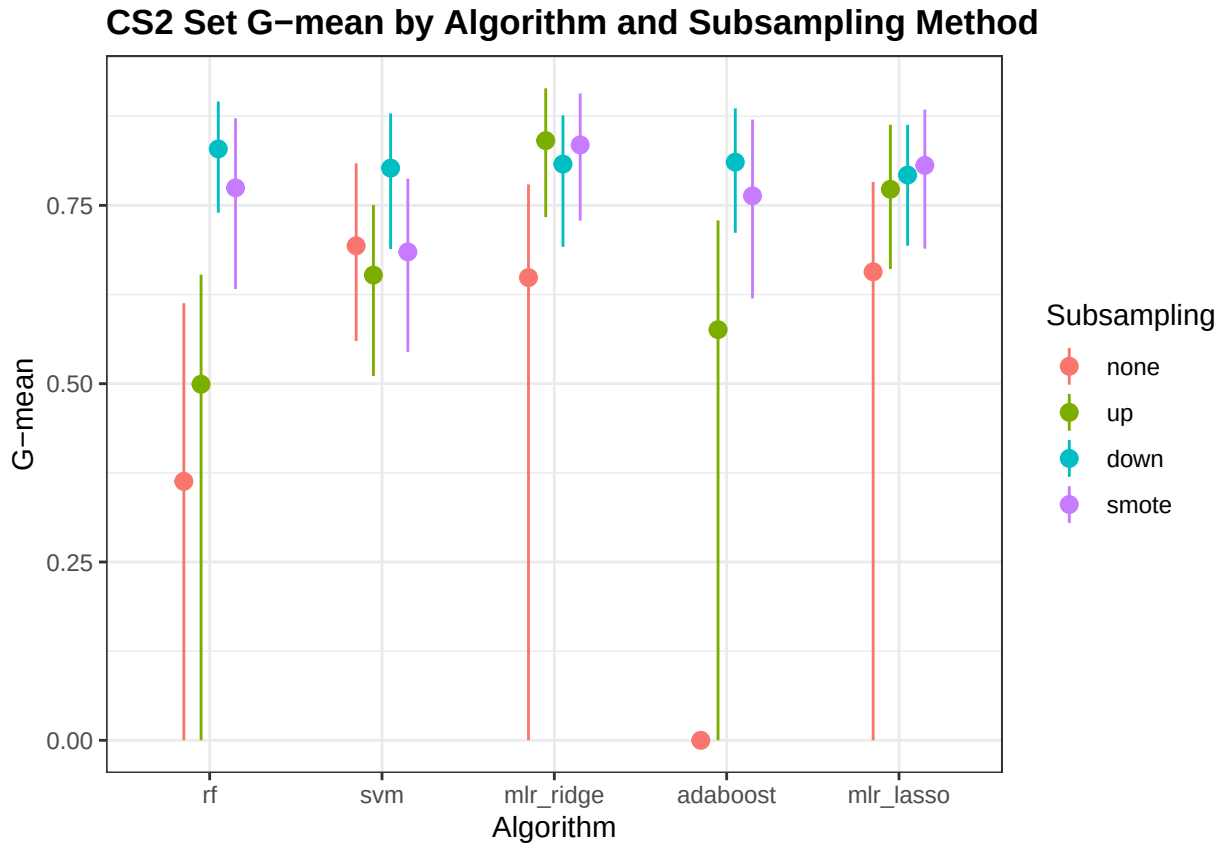


Figure 4.39: CS2 Set G-mean

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

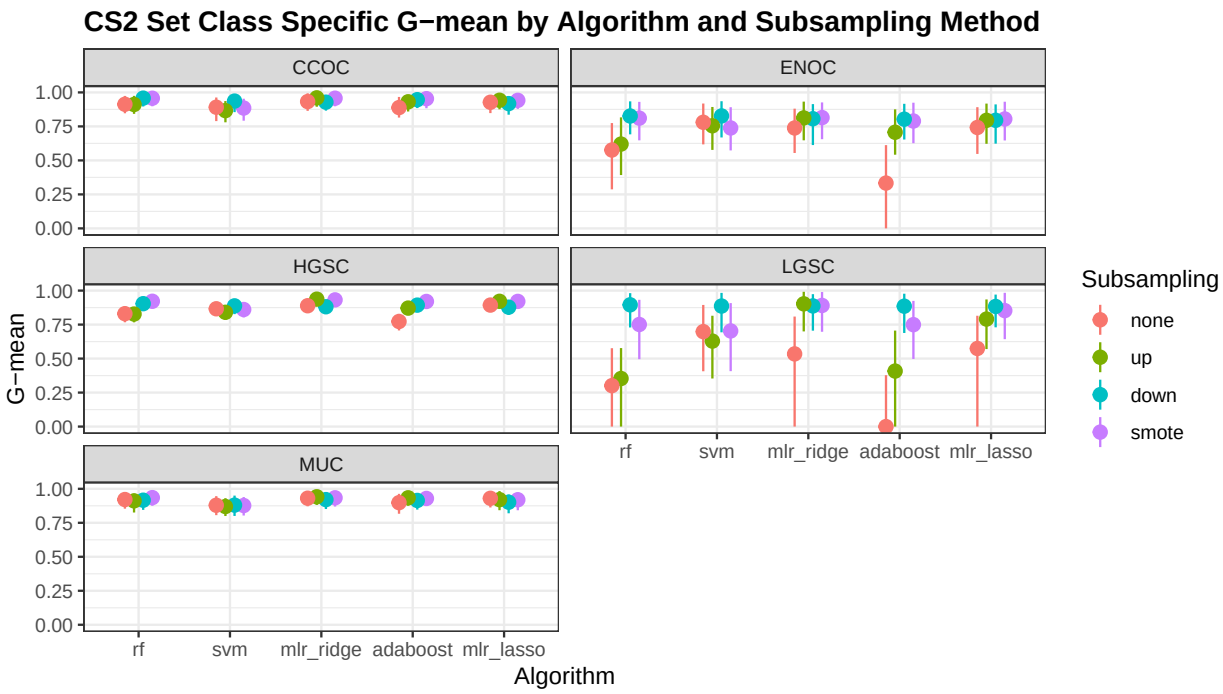


Figure 4.40: CS2 Set Class-Specific G-mean

Table 4.39: SMOTE Kappa by Algorithm and Dataset

| dataset | rf | svm | mlr_ridge | adaboost | mlr_lasso |
|---|---|---|---|---|---|
| Training | **0.843** | 0.816 | 0.782 | 0.821 | 0.769 |
| CS1 | 0.777 | 0.768 | 0.767 | 0.766 | 0.746 |
| CS2 | 0.802 | 0.766 | 0.781 | 0.797 | 0.75 |

## 4.5 SMOTE Kappa Summary

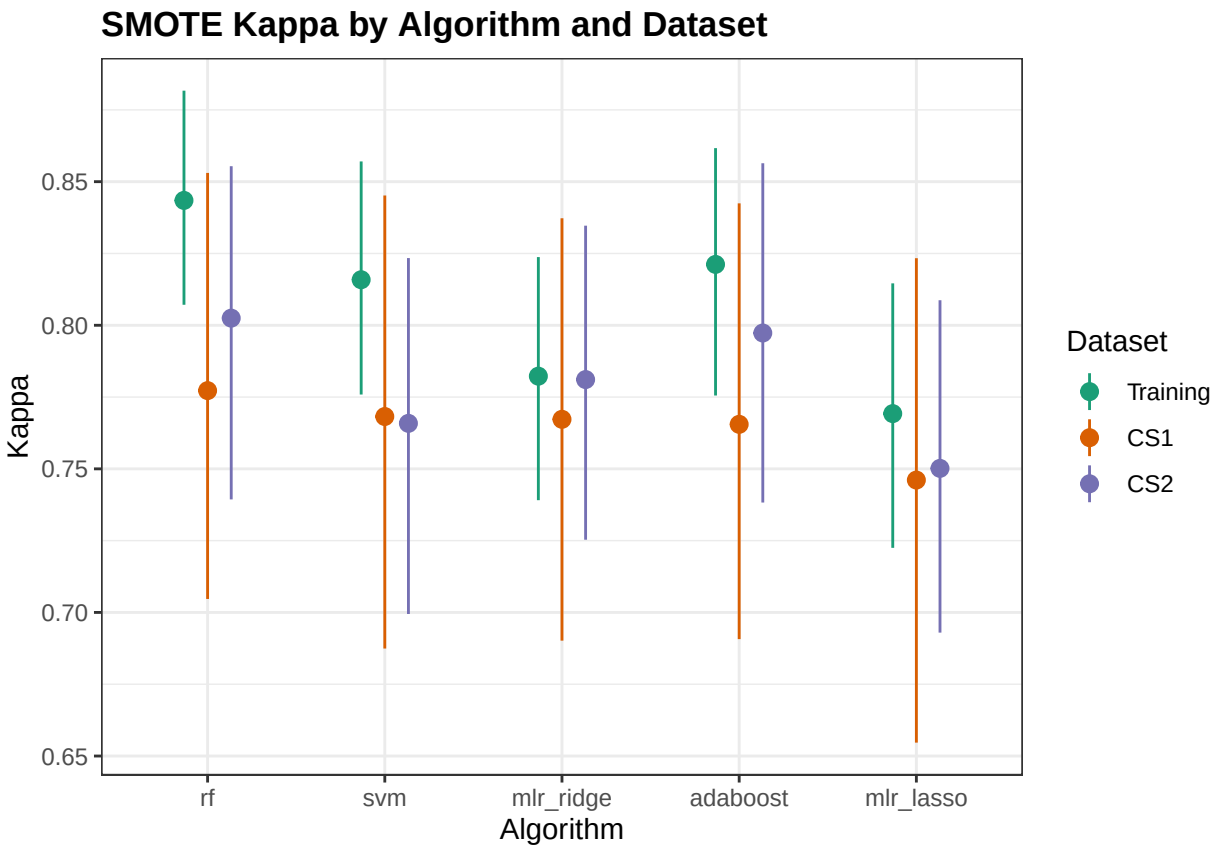


Figure 4.41: SMOTE Kappa by Algorithm and Dataset

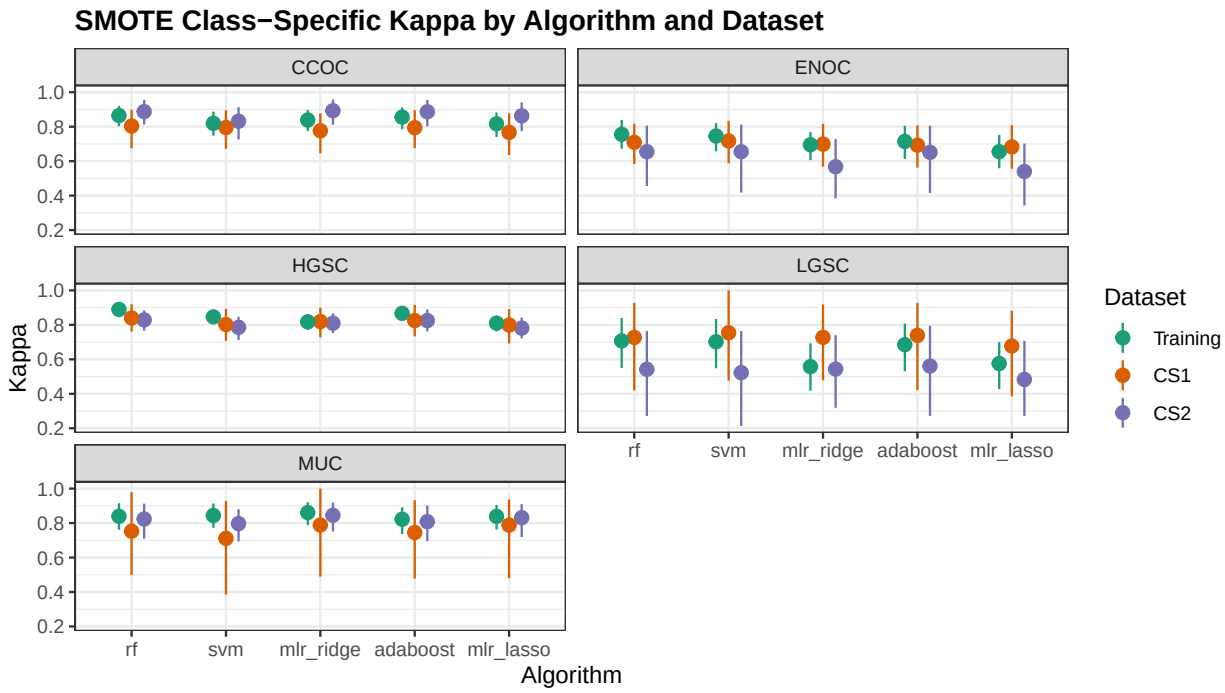


Figure 4.42: SMOTE Class-Specific Kappa by Algorithm and Dataset

## 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 | 0.927 | 0.839 | 0.975 | 0.875 | 0.892 |
| sequential-up | 0.911 | 0.97 | 0.973 | 0.929 | 0.889 |
| sequential-none | 0.909 | 0.972 | 0.973 | 0.933 | 0.897 |
| sequential-down | 0.905 | 0.971 | 0.964 | 0.933 | 0.897 |
| sequential-smote | 0.907 | 0.97 | 0.973 | 0.933 | 0.897 |
| 2S-rf-up | 0.925 | 0.844 | 0.975 | 0.875 | 0.896 |
| 2S-svm-up | 0.894 | 0.853 | 0.973 | 0.914 | 0.885 |
| 2S-svm-smote | 0.889 | 0.852 | 0.972 | 0.909 | 0.88 |
| 2S-rf-smote | 0.919 | 0.839 | 0.974 | 0.875 | 0.889 |
| 2S-mlr_ridge-up | 0.892 | 0.833 | 0.962 | 0.87 | 0.909 |
| 2S-adaboost-up | 0.909 | 0.831 | 0.974 | 0.871 | 0.882 |
| 2S-svm-none | 0.894 | 0.853 | 0.973 | 0.917 | 0.885 |
| 2S-mlr_ridge-none | 0.899 | 0.833 | 0.968 | 0.889 | 0.912 |
| 2S-mlr_ridge-smote | 0.894 | 0.833 | 0.966 | 0.867 | 0.909 |
| 2S-svm-down | 0.88 | 0.829 | 0.963 | 0.909 | 0.875 |
| 2S-adaboost-down | 0.9 | 0.806 | 0.964 | 0.867 | 0.871 |
| 2S-adaboost-smote | 0.911 | 0.829 | 0.971 | 0.88 | 0.871 |
| 2S-adaboost-none | 0.914 | 0.825 | 0.971 | 0.889 | 0.887 |
| adaboost-up | 0.873 | 0.765 | 0.976 | 0.667 | 0.857 |
| 2S-mlr_lasso-none | 0.9 | 0.825 | 0.967 | 0.867 | 0.896 |
| 2S-mlr_ridge-down | 0.892 | 0.828 | 0.959 | 0.857 | 0.9 |
| 2S-rf-down | 0.909 | 0.824 | 0.964 | 0.867 | 0.871 |
| 2S-mlr_lasso-smote | 0.901 | 0.825 | 0.965 | 0.865 | 0.883 |
| rf-none | 0.875 | 0.733 | 0.969 | 0.375 | 0.873 |
| rf-smote | 0.873 | 0.769 | 0.977 | 0.714 | 0.849 |
| svm-none | 0.857 | 0.776 | 0.971 | 0.667 | 0.857 |
| mlr_ridge-none | 0.871 | 0.758 | 0.97 | 0.37 | 0.885 |
| rf-up | 0.873 | 0.776 | 0.971 | 0.526 | 0.877 |
| adaboost-smote | 0.865 | 0.732 | 0.972 | 0.692 | 0.833 |
| adaboost-none | 0.862 | 0.702 | 0.961 | 0.167 | 0.857 |
| 2S-mlr_lasso-down | 0.897 | 0.808 | 0.956 | 0.839 | 0.857 |
| mlr_ridge-smote | 0.85 | 0.716 | 0.959 | 0.571 | 0.868 |
| svm-smote | 0.831 | 0.762 | 0.969 | 0.71 | 0.851 |
| mlr_lasso-none | 0.865 | 0.734 | 0.971 | 0.483 | 0.862 |
| mlr_ridge-up | 0.846 | 0.675 | 0.948 | 0.545 | 0.862 |
| mlr_lasso-smote | 0.83 | 0.676 | 0.957 | 0.588 | 0.847 |
| 2S-mlr_lasso-up | 0.897 | 0.825 | 0.96 | 0.857 | 0.892 |
| svm-up | 0.836 | 0.754 | 0.971 | 0.72 | 0.846 |
| mlr_ridge-down | 0.829 | 0.625 | 0.92 | 0.418 | 0.841 |
| rf-down | 0.849 | 0.645 | 0.934 | 0.481 | 0.8 |
| adaboost-down | 0.837 | 0.617 | 0.929 | 0.473 | 0.771 |
| mlr_lasso-up | 0.8 | 0.646 | 0.958 | 0.615 | 0.821 |
| svm-down | 0.781 | 0.639 | 0.931 | 0.519 | 0.8 |
| mlr_lasso-down | 0.805 | 0.602 | 0.907 | 0.377 | 0.785 |

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 | 0.874 | 0.772 | 0.419 | 0.745 |
| sequential-none | 0.902 | 0.910 | 0.917 | 0.485 | 0.711 |
| 2S-rf-smote | 0.903 | 0.904 | 0.784 | 0.400 | 0.667 |
| sequential-smote | 0.907 | 0.923 | 0.921 | 0.556 | 0.800 |
| rf-none | 0.899 | 0.881 | 0.533 | 0.133 | 0.760 |

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-up**: sequential algorithm with upsampling. The sequence of models and algorithms used are:
  - HGSC vs. non-HGSC using adaboost
  - MUC vs. non-MUC using random forest
  - CCOC vs. non-CCOC using random forest
  - ENOC vs. LGSC using random forest
- **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
- **sequential-down**: sequential algorithm with downsampling. 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 ridge regression
  - 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.

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

| method | HGSC | CCOC | ENOC | LGSC | MUC |
|---|---|---|---|---|---|
| 2S-rf-none | 0.938 | 0.946 | 0.899 | 0.811 | 0.764 |
| sequential-none | 0.948 | 0.963 | 0.963 | 0.789 | 0.784 |
| 2S-rf-smote | 0.933 | 0.940 | 0.882 | 0.750 | 0.778 |
| sequential-smote | 0.940 | 0.951 | 0.967 | 0.821 | 0.821 |
| rf-none | 0.929 | 0.857 | 0.630 | 0.182 | 0.727 |

### 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.