

Ovarian Cancer Histotypes: Report of Statistical Findings

Derek Chiu

2021-07-27

Contents

| | |
|---|-----------|
| Preface | 8 |
| 1 Introduction | 9 |
| 2 Methods | 10 |
| 2.1 Data Processing | 10 |
| 2.2 Normalization Between CodeSets | 14 |
| 2.3 Histotype Classification | 15 |
| 3 Validation | 17 |
| 3.1 Full Data Distributions | 17 |
| 3.2 Training Set Distributions | 17 |
| 3.3 Normalization | 21 |
| 3.4 Common Sample Distributions | 55 |
| 3.5 Histotype Distribution in Classifier Datasets | 55 |
| 4 Results | 57 |
| 4.1 Training Set | 57 |
| 4.2 Two-Step Training Set | 66 |
| 4.3 CS1 Set | 82 |
| 4.4 CS2 Set | 90 |
| 4.5 SMOTE Kappa Summary | 98 |
| 4.6 Overlap with SPOT | 99 |

List of Tables

| | |
|---|----|
| 2.1 Cohort Distribution amongst CodeSets | 14 |
| 2.2 Distinct Cohort Distribution amongst CodeSets | 15 |
| 3.1 All CodeSet Histotype Groups | 17 |
| 3.2 All CodeSet Histotypes | 18 |
| 3.3 Common Summary ID CodeSet Histotypes | 18 |
| 3.4 All CodeSet Major Histotypes | 18 |
| 3.5 CS1 Histotypes | 19 |
| 3.6 CS2 Histotypes | 19 |
| 3.7 CS3 Histotypes | 20 |
| 3.8 CS1 Training Set Histotypes | 20 |
| 3.9 CS2 Training Set Histotypes | 20 |
| 3.10 Random1 CS1 vs. CS3 Median Concordance Measures by Histotypes | 28 |
| 3.11 Random1 CS2 vs. CS3 Median Concordance Measures by Histotypes | 28 |
| 3.12 CS2 vs CS3 Random1 Normalized 100 Runs of Summary Concordance Measures | 29 |
| 3.13 Random3 HGSC CS1 vs. CS3 Median Concordance Measures by Histotypes | 30 |
| 3.14 Random3 HGSC CS2 vs. CS3 Median Concordance Measures by Histotypes | 30 |
| 3.15 Pools Non-Normalized CS2 vs. CS3 Median Concordance Measures by Histotypes | 33 |
| 3.16 Pools Normalized CS2 vs. CS3 Median Concordance Measures by Histotypes | 33 |
| 3.17 Random3 Samples Comparisons Statistics by Histotypes | 38 |
| 3.18 Random2 Samples Comparisons Statistics by Histotypes | 39 |
| 3.19 Random1 Samples Comparisons Statistics by Histotypes | 40 |
| 3.20 DSC for CS2 vs CS3 Comparisons | 40 |
| 3.21 DSC for CS1 vs CS3 Comparisons | 44 |
| 3.22 DSC for CS3 vs CS5 Set B/A Comparisons | 48 |
| 3.23 DSC for CS3 vs CS5 Set C/A Comparisons | 52 |
| 3.24 All Common Samples Histotype Distribution | 53 |
| 3.25 Distinct Common Samples Histotype Distribution | 53 |
| 3.26 Distinct Common CS2 and CS3 Samples Histotype Distribution | 54 |

| | | |
|------|---|----|
| 3.27 | Common Samples Across Sites Histotype Distribution | 54 |
| 3.28 | Distinct Common Samples Across Sites Histotype Distribution | 54 |
| 3.29 | CS3/CS4/CS5 Common Samples Histotype Distribution | 54 |
| 3.30 | CS3/CS4/CS5 Pools Distribution | 54 |
| 3.31 | Full Training Set Histotype Distribution | 55 |
| 3.32 | Full Training Set Histotype Distribution by CodeSet | 55 |
| 3.33 | CS1 All Training Set Histotype Distribution | 55 |
| 3.34 | CS2 All Training Set Histotype Distribution | 56 |
| 3.35 | Confirmation Set Histotype Distribution | 56 |
| 3.36 | Validation Set Histotype Distribution | 56 |
| 4.1 | Training Set Accuracy by Algorithm and Subsampling Method | 58 |
| 4.2 | Training Set Class-Specific Accuracy by Algorithm and Subsampling Method | 59 |
| 4.3 | Training Set Macro-Averaged F1-Score by Algorithm and Subsampling Method | 60 |
| 4.4 | Training Set Class-Specific F1-Score by Algorithm and Subsampling Method | 61 |
| 4.5 | Training Set Kappa by Algorithm and Subsampling Method | 62 |
| 4.6 | Training Set Class-Specific Kappa by Algorithm and Subsampling Method | 63 |
| 4.7 | Training Set G-mean by Algorithm and Subsampling Method | 64 |
| 4.8 | Training Set Class-Specific G-mean by Algorithm and Subsampling Method | 65 |
| 4.9 | Training Set Step 1 Accuracy by Algorithm and Subsampling Method | 67 |
| 4.10 | Training Set Step 2 Accuracy by Algorithm and Subsampling Method | 67 |
| 4.11 | Training Set Step 1 Class-Specific Accuracy by Algorithm and Subsampling Method | 69 |
| 4.12 | Training Set Step 2 Class-Specific Accuracy by Algorithm and Subsampling Method | 69 |
| 4.13 | Training Set Step 1 Macro-Averaged F1-Score by Algorithm and Subsampling Method | 71 |
| 4.14 | Training Set Step 2 Macro-Averaged F1-Score by Algorithm and Subsampling Method | 71 |
| 4.15 | Training Set Step 1 Class-Specific F1-Score by Algorithm and Subsampling Method | 73 |
| 4.16 | Training Set Step 2 Class-Specific F1-Score by Algorithm and Subsampling Method | 73 |
| 4.17 | Training Set Step 1 Kappa by Algorithm and Subsampling Method | 75 |
| 4.18 | Training Set Step 2 Kappa by Algorithm and Subsampling Method | 75 |
| 4.19 | Training Set Step 1 Class-Specific Kappa by Algorithm and Subsampling Method | 77 |
| 4.20 | Training Set Step 2 Class-Specific Kappa by Algorithm and Subsampling Method | 77 |
| 4.21 | Training Set Step 1 G-mean by Algorithm and Subsampling Method | 79 |
| 4.22 | Training Set Step 2 G-mean by Algorithm and Subsampling Method | 79 |
| 4.23 | Training Set Step 1 Class-Specific G-mean by Algorithm and Subsampling Method | 81 |
| 4.24 | Training Set Step 2 Class-Specific G-mean by Algorithm and Subsampling Method | 81 |
| 4.25 | CS1 Set Accuracy by Algorithm and Subsampling Method | 82 |
| 4.26 | CS1 Set Class-Specific Accuracy by Algorithm and Subsampling Method | 83 |

| | |
|--|----|
| 4.27 CS1 Set Macro-Averaged F1-Score by Algorithm and Subsampling Method | 84 |
| 4.28 CS1 Set Class-Specific F1-Score by Algorithm and Subsampling Method | 85 |
| 4.29 CS1 Set Kappa by Algorithm and Subsampling Method | 86 |
| 4.30 CS1 Set Class-Specific Kappa by Algorithm and Subsampling Method | 87 |
| 4.31 CS1 Set G-mean by Algorithm and Subsampling Method | 88 |
| 4.32 CS1 Set Class-Specific G-mean by Algorithm and Subsampling Method | 89 |
| 4.33 CS2 Set Accuracy by Algorithm and Subsampling Method | 90 |
| 4.34 CS2 Set Class-Specific Accuracy by Algorithm and Subsampling Method | 91 |
| 4.35 CS2 Set Macro-Averaged F1-Score by Algorithm and Subsampling Method | 92 |
| 4.36 CS2 Set Class-Specific F1-Score by Algorithm and Subsampling Method | 93 |
| 4.37 CS2 Set Kappa by Algorithm and Subsampling Method | 94 |
| 4.38 CS2 Set Class-Specific Kappa by Algorithm and Subsampling Method | 95 |
| 4.39 CS2 Set G-mean by Algorithm and Subsampling Method | 96 |
| 4.40 CS2 Set Class-Specific G-mean by Algorithm and Subsampling Method | 97 |
| 4.41 SMOTE Kappa by Algorithm and Dataset | 98 |

List of Figures

| | | |
|------|--|----|
| 3.1 | Random3 Non-Normalized Concordance Measure Distributions | 23 |
| 3.2 | Random3 Normalized Concordance Measure Distributions | 24 |
| 3.3 | Random2 Non-Normalized Concordance Measure Distributions | 25 |
| 3.4 | Random2 Normalized Concordance Measure Distributions | 26 |
| 3.5 | Random1 Non-Normalized Concordance Measure Distributions | 27 |
| 3.6 | Random1 Normalized Concordance Measure Distributions | 28 |
| 3.7 | CS2 vs. CS3 Random1 Normalized 100 Runs of Median Concordance Measures | 29 |
| 3.8 | Random3 HGSC Normalized Concordance Measure Distributions | 30 |
| 3.9 | Cross-Site Random1 Non-Normalized Concordance Measure Distributions | 31 |
| 3.10 | CS2Non vs. CS2Pools Concordance Measure Distributions | 32 |
| 3.11 | CS2 Non-Normalized Pools vs. CS3 Concordance Measure Distributions | 33 |
| 3.12 | CS2 Normalized Pools vs. CS3 Concordance Measure Distributions | 33 |
| 3.13 | USC-Non vs. USC-Pools Concordance Measure Distributions | 34 |
| 3.14 | USC-Non vs. VAN-Non Concordance Measure Distributions | 34 |
| 3.15 | USC-Pools vs. VAN-Non Concordance Measure Distributions | 34 |
| 3.16 | USC vs. VAN Comparisons of Concordance Measure Distributions | 35 |
| 3.17 | AOC-Non vs. AOC-Pools Concordance Measure Distributions | 35 |
| 3.18 | AOC-Non vs. VAN-Non Concordance Measure Distributions | 36 |
| 3.19 | AOC-Pools vs. VAN-Non Concordance Measure Distributions | 36 |
| 3.20 | AOC vs. VAN Comparisons of Concordance Measure Distributions | 37 |
| 3.21 | Random3 Samples Comparisons of Concordance Measure Distributions | 38 |
| 3.22 | Random2 Samples Comparisons of Concordance Measure Distributions | 39 |
| 3.23 | Random1 Samples Comparisons of Concordance Measure Distributions | 40 |
| 3.24 | Random1 Concordance Measure Distributions | 41 |
| 3.25 | Random1 + Pools Concordance Measure Distributions | 42 |
| 3.26 | CS1 CodeSet Chaining Concordance Measure Distributions | 43 |
| 3.27 | CS1 CodeSet Chaining Concordance Measure Distributions 2 | 44 |
| 3.28 | Pairwise Genes by Top/Bottom 3 Rc for CS1 vs. CS3 | 45 |

| | |
|--|----|
| 3.29 CS2 CodeSet Chaining Concordance Measure Distributions | 46 |
| 3.30 CS5 Set B/A Chaining Concordance Measure Distributions | 47 |
| 3.31 CS5 Set B/A Chaining Concordance Measure Distributions 2 | 48 |
| 3.32 CS4 Set A Chaining Concordance Measure Distributions | 49 |
| 3.33 CS4 and CS5 using Set B Concordance Measure Distributions | 50 |
| 3.34 CS5 Set C/A Chaining Concordance Measure Distributions | 51 |
| 3.35 CS5 Set C/A Chaining Concordance Measure Distributions 2 | 52 |
| 3.36 CS4 and CS5 using Set C Concordance Measure Distributions | 53 |
| | |
| 4.1 Training Set Accuracy | 57 |
| 4.2 Training Set Class-Specific Accuracy | 58 |
| 4.3 Training Set F1-Score | 60 |
| 4.4 Training Set Class-Specific F1-Score | 61 |
| 4.5 Training Set Kappa | 62 |
| 4.6 Training Set Class-Specific Kappa | 63 |
| 4.7 Training Set G-mean | 64 |
| 4.8 Training Set Class-Specific G-mean | 65 |
| 4.9 Training Set Step 1 Accuracy | 66 |
| 4.10 Training Set Step 2 Accuracy | 67 |
| 4.11 Training Set Step 1 Class-Specific Accuracy | 68 |
| 4.12 Training Set Step 2 Class-Specific Accuracy | 68 |
| 4.13 Training Set Step 1 F1-Score | 70 |
| 4.14 Training Set Step 2 F1-Score | 71 |
| 4.15 Training Set Step 1 Class-Specific F1-Score | 72 |
| 4.16 Training Set Step 2 Class-Specific F1-Score | 72 |
| 4.17 Training Set Step 1 Kappa | 74 |
| 4.18 Training Set Step 2 Kappa | 75 |
| 4.19 Training Set Step 1 Class-Specific Kappa | 76 |
| 4.20 Training Set Step 2 Class-Specific Kappa | 76 |
| 4.21 Training Set Step 1 G-mean | 78 |
| 4.22 Training Set Step 2 G-mean | 79 |
| 4.23 Training Set Step 1 Class-Specific G-mean | 80 |
| 4.24 Training Set Step 2 Class-Specific G-mean | 80 |
| 4.25 CS1 Set Accuracy | 82 |
| 4.26 CS1 Set Class-Specific Accuracy | 83 |
| 4.27 CS1 Set F1-Score | 84 |
| 4.28 CS1 Set Class-Specific F1-Score | 85 |

| | | |
|------|---|----|
| 4.29 | CS1 Set Kappa | 86 |
| 4.30 | CS1 Set Class-Specific Kappa | 87 |
| 4.31 | CS1 Set G-mean | 88 |
| 4.32 | CS1 Set Class-Specific G-mean | 89 |
| 4.33 | CS2 Set Accuracy | 90 |
| 4.34 | CS2 Set Class-Specific Accuracy | 91 |
| 4.35 | CS2 Set F1-Score | 92 |
| 4.36 | CS2 Set Class-Specific F1-Score | 93 |
| 4.37 | CS2 Set Kappa | 94 |
| 4.38 | CS2 Set Class-Specific Kappa | 95 |
| 4.39 | CS2 Set G-mean | 96 |
| 4.40 | CS2 Set Class-Specific G-mean | 97 |
| 4.41 | SMOTE Kappa by Algorithm and Dataset | 98 |
| 4.42 | SMOTE Class-Specific Kappa by Algorithm and Dataset | 99 |

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.

In the NanoString CodeSets, we also run into a problem with trying to find suitable control pools to normalize the gene expression. For prospective NanoString runs, the pools can be specifically chosen, but for retrospective runs, we have to utilize a combination of common samples and common genes as references for normalization.

The supervised learning is performed under a consensus framework: we consider various classification algorithms and use evaluation metrics to help make decisions of which methods to carry forward for downstream analysis.

2. Methods

2.1 Data Processing

RNA was extracted from FFPE ovarian carcinoma samples and expression was quantified using NanoString nCounter. Samples were run in three CodeSets. Some samples or pools of samples were repeated across CodeSets for expression normalization. Normalizing CS2 to CS3 can easily follow the [PrOType](#) method for HGSC subtypes because both CodeSets have pool samples. A different technique is implemented when normalizing across CS1, CS2, and CS3 where we use common samples and genes as reference sets.

2.1.1 Raw Data

NanoString CodeSets contained a mix of all probes of interest, six positive controls spiked-in at fixed proportional concentrations (0.125- 128 fM), and eight negative controls (probes without a corresponding target). Gene targets also included 5 housekeeping genes: POLR1B, SDHA, PGK1, ACTB, RPL19. Gene selection was made from top ranked differential gene expression analysis between ovarian cancer histotypes and molecular subtypes of HGSC, as well as containing some genes of interest from unrelated projects. Gene targets in each subsequent CodeSet were re-curated, where non-informative genes were dropped and new potential differentiating genes were added.

There are 3 NanoString CodeSets:

- CS1: OvCa2103_C953
 - Samples = 412
 - Genes = 275
- CS2: PrOTYPE2_v2_C1645
 - Samples = 1223
 - Genes = 384
- CS3: OTTA2014_C2822
 - Samples = 5424
 - Genes = 532

These datasets contain raw counts extracted straight from NanoString RCC files.

2.1.2 Housekeeping Genes

The first normalization step is to normalize all endogenous genes to housekeeping genes (POLR1B, SDHA, PGK1, ACTB, RPL19; reference genes expressed in all cells). We normalize by subtracting the average \log_2 housekeeping gene expression from the \log_2 endogenous gene expression:

$$\log_2 \text{ endogenous expression} - \log_2 \text{ average housekeeping expression} = \text{relative expression}$$

The updated CodeSet dimensions are now:

- CS1: OvCa2103_C953
 - Samples = 412
 - Genes = 256
- CS2: PrOTYPE2_v2_C1645
 - Samples = 1223
 - Genes = 365
- CS3: OTTA2014_C2822
 - Samples = 5424
 - Genes = 513

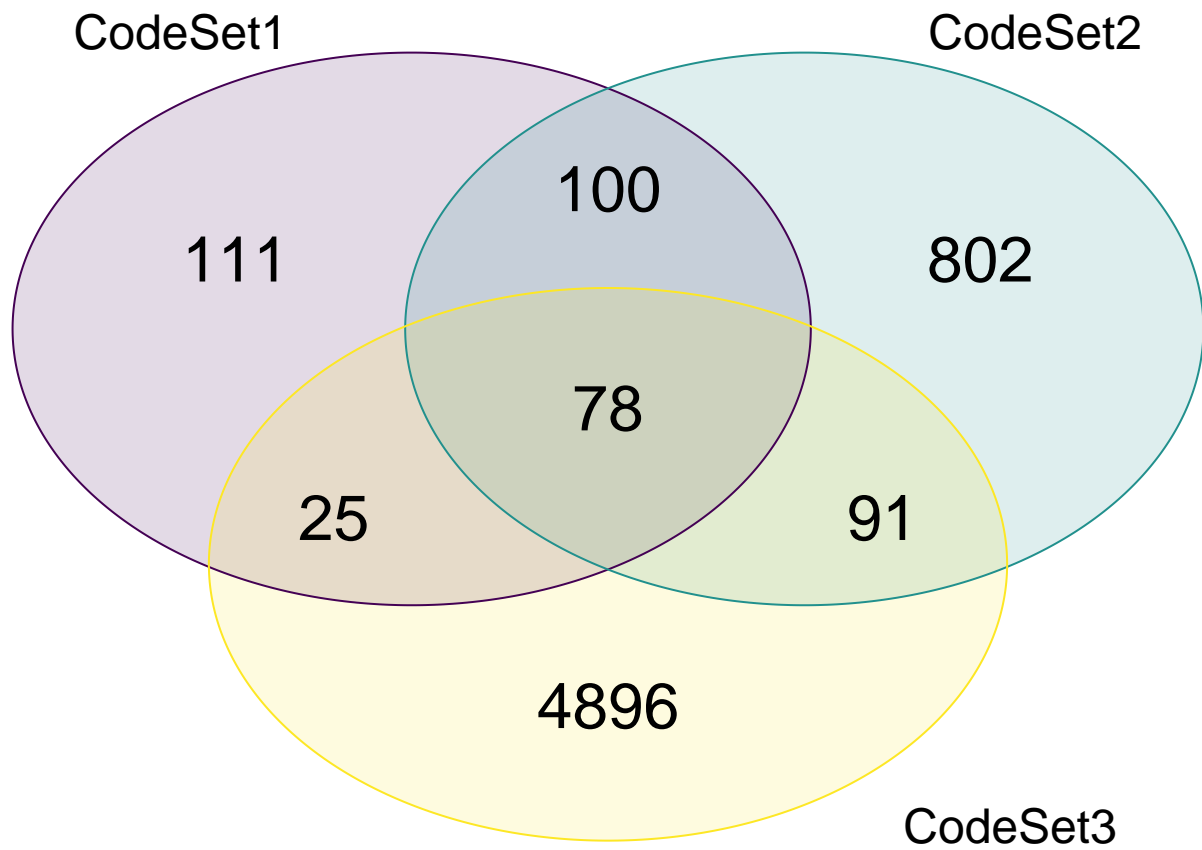
The number of genes are reduced by 19: 5 housekeeping, 8 negative, 6 positive (the latter 2 types are not used).

2.1.3 Common Samples and Genes

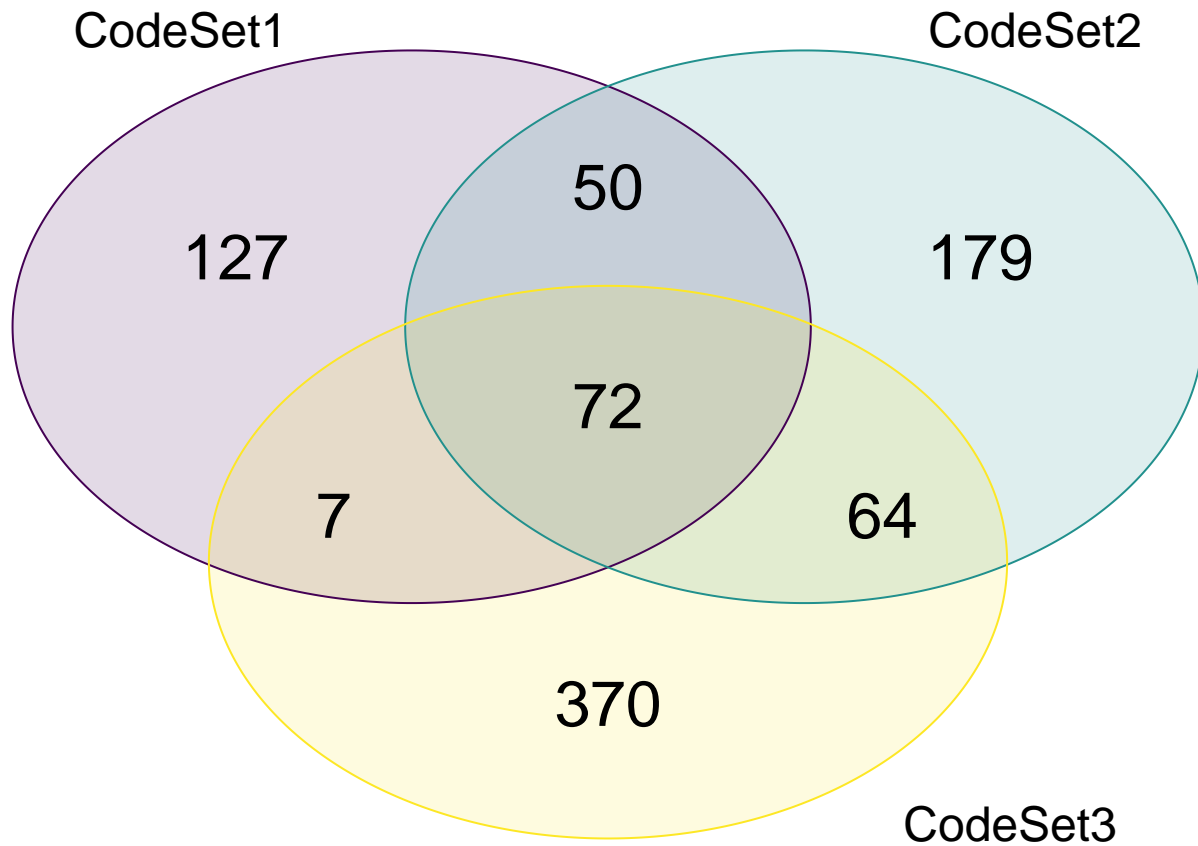
Since the reference pool samples only exist in CS2 and CS3, we need to find an alternative method to normalize all three CodeSets. One method is to select common samples and common genes that exist in all three. We found 72 common genes. Using the `summaryID` identifier, we also found 78 common summary IDs, translating to 320 samples. The number of samples that were matched to each CodeSet differed:

- CS1: OvCa2103_C953
 - Samples = 93
 - Genes = 72
- CS2: PrOTYPE2_v2_C1645
 - Samples = 87
 - Genes = 72
- CS3: OTTA2014_C2822
 - Samples = 140
 - Genes = 72

2.1.3.1 Overlap of common samples by summary ID



2.1.3.2 Overlap of common genes



*Excluding housekeeping genes and controls

2.1.4 CS1 Training Set Generation

We use the reference method to normalize CS1 to CS3.

- CS1 reference set: duplicate samples from CS1
 - Samples = 25
 - Genes = 72
- CS3 reference set: corresponding samples in CS3 also found in CS1 reference set
 - Samples = 20
 - Genes = 72
- CS1 validation set: remaining CS1 samples with reference set removed
 - Samples = 387
 - Genes = 72

The final CS1 training set has 304 samples on 72 genes after normalization and keeping only the major histotypes of interest.

Table 2.1: Cohort Distribution amongst CodeSets

| cohort | cs1 | cs2 | cs3 |
|-----------|-----|-----|------|
| MAYO | 6 | 63 | NA |
| MTL | 3 | 59 | NA |
| OOU | 108 | 43 | 19 |
| OOUE | 32 | 30 | 11 |
| VOA | 145 | 122 | 538 |
| ICON7 | NA | 416 | NA |
| JAPAN | NA | 8 | NA |
| OVAR3 | NA | 150 | NA |
| POOL-CTRL | NA | 12 | NA |
| DOVE4 | NA | NA | 1160 |
| POOL-1 | NA | NA | 31 |
| POOL-2 | NA | NA | 14 |
| POOL-3 | NA | NA | 13 |
| TNCO | NA | NA | 691 |

2.1.5 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 = 1214
 - Genes = 365

The final CS2 training set has 945 samples on 136 (common) genes after normalization and keeping only the major histotypes of interest.

2.1.6 Cohort Distribution

CodeSets comprised samples from sites collected internationally as shown below. Note that the CS3 pools sample total (n=58) shown here include those that are not used as reference pools, following previous normalization methods. In particular, the distribution of CS3 pools actually used for normalization (n=22) is POOL1 = 12, POOL2 = 5, POOL3 = 5.

2.2 Normalization Between CodeSets

After normalization to housekeeping genes and filtering for the five major histotypes of interest, as determined by pathology review and/or IHC, two methods were used to normalize data between CodeSets.

Table 2.2: Distinct Cohort Distribution amongst CodeSets

| cohort | cs1 | cs2 | cs3 |
|-----------|-----|-----|------|
| MAYO | 6 | 62 | NA |
| MTL | 3 | 59 | NA |
| OOU | 99 | 43 | 19 |
| OOUE | 31 | 30 | 11 |
| VOA | 136 | 107 | 452 |
| ICON7 | NA | 383 | NA |
| JAPAN | NA | 8 | NA |
| OVAR3 | NA | 150 | NA |
| POOL-CTRL | NA | 3 | NA |
| DOVE4 | NA | NA | 1094 |
| POOL-1 | NA | NA | 12 |
| POOL-2 | NA | NA | 5 |
| POOL-3 | NA | NA | 5 |
| TNCO | NA | NA | 674 |

2.2.1 Common Samples Method

The common samples method was used to normalize CodeSet1, 2, and 3, where common samples and genes were used as reference sets. Among the samples repeated in all CodeSets we normalized using either: a random set of 3 samples from each major histotype (random3; n=15), a random set of 2 samples from each major histotype (random2; n=10), or a random set of 1 sample from each major histotype (random1; n=5). In each case CodeSet3 expression (X_3) was held fixed, while CodeSet1/2 expression (X_1 and X_2) were normalized to CodeSet3 by subtracting the average gene expression from the CodeSet1/2 reference set (R_1 or R_2) and adding the average gene expression of the CodeSet3 reference set (R_3). Alternatively, X_1 (norm) = $X_1 - R_1 + R_3$ would calibrate CodeSet1 to CodeSet3.

2.2.2 Pools Method

The pools method was used to normalize CodeSet2 and CodeSet3. The three reference pools, regularly assayed mixes of samples representing all histotypes, were run in CodeSet2 and CodeSet3 only. CodeSet2 contained 12 reference pool samples (Pool 1 = 4, Pool 2 = 4, Pool 3 = 4) and CodeSet3 contained 22 reference pool samples (Pool 1 = 12, Pool 2 = 5, Pool 3 = 5). Similar to the common samples method, CodeSet2 was normalized to CodeSet3 via: X_2 (norm) = $X_2 - R_2 + R_3$ where R is the average expression of the reference pool samples in the respective CodeSet. This method of pool normalization was also used by PrOType to classify HGSC subtypes

2.2.3 Concordance Comparison

Concordance between CodeSets using the different normalization strategies was compared in common samples, excluding those used for the normalization, using Pearson’s correlation coefficient (R^2), coefficient of accuracy (Ca), and Lin’s concordance correlation ($R_c = R^2 \times Ca$).

2.3 Histotype Classification

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. Validation

3.1 Full Data Distributions

The histotype distributions on the full data are shown below.

3.2 Training Set Distributions

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

Table 3.1: All CodeSet Histotype Groups

| hist_gr | CS1 | CS2 | CS3 |
|----------|-----|-----|------|
| HGSC | 169 | 757 | 2453 |
| non-HGSC | 196 | 373 | 677 |

Table 3.2: All CodeSet Histotypes

| revHist | CS1 | CS2 | CS3 |
|---|-----|-----|------|
| CARCINOMA-NOS | 0 | 61 | 23 |
| Carcinoma, NOS | 0 | 0 | 2 |
| CCOC | 57 | 68 | 182 |
| CCOC-MCT | 0 | 1 | 0 |
| Cell-Line | 17 | 48 | 13 |
| CTRL | 0 | 12 | 0 |
| ENOC | 61 | 30 | 272 |
| ENOC-CCOC | 0 | 7 | 0 |
| ERROR | 0 | 3 | 0 |
| HGSC | 169 | 757 | 2453 |
| HGSC-MCT | 0 | 1 | 0 |
| LGSC | 22 | 29 | 50 |
| MBOT | 0 | 20 | 3 |
| MET-NOP | 0 | 21 | 0 |
| MIXED (ENOC/CCOC) | 0 | 0 | 1 |
| MIXED (ENOC/LGSC) | 0 | 0 | 1 |
| MIXED (HGSC/CCOC) | 0 | 0 | 1 |
| mixed cell | 0 | 0 | 7 |
| MMMT | 0 | 0 | 30 |
| MUC | 20 | 61 | 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 | 10 | 3 |
| Serous | 0 | 0 | 2 |
| serous LMP | 0 | 0 | 1 |
| SQAMOUS | 0 | 1 | 0 |

Table 3.3: Common Summary ID CodeSet Histotypes

| revHist | CS1 | CS2 | CS3 |
|-----------|-----|-----|-----|
| CCOC | 3 | 4 | 9 |
| Cell-Line | 4 | 5 | 5 |
| ENOC | 4 | 4 | 9 |
| HGSC | 68 | 64 | 98 |
| LGSC | 7 | 5 | 8 |
| MUC | 7 | 5 | 11 |

Table 3.4: All CodeSet Major Histotypes

| revHist | CS1 | CS2 | CS3 | CS1_percent | CS2_percent | CS3_percent |
|---------|-----|-----|------|-------------|-------------|-------------|
| CCOC | 57 | 68 | 182 | 17.3 | 7.2 | 6.0 |
| ENOC | 61 | 30 | 272 | 18.5 | 3.2 | 9.0 |
| HGSC | 169 | 757 | 2453 | 51.4 | 80.1 | 80.9 |
| LGSC | 22 | 29 | 50 | 6.7 | 3.1 | 1.6 |
| MUC | 20 | 61 | 77 | 6.1 | 6.5 | 2.5 |

Table 3.5: CS1 Histotypes

| CodeSet | revHist | n |
|---------|-----------|-----|
| CS1 | CCOC | 57 |
| CS1 | Cell-Line | 17 |
| CS1 | ENOC | 61 |
| CS1 | HGSC | 169 |
| CS1 | LGSC | 22 |
| CS1 | MUC | 20 |
| CS1 | SBOT | 19 |

Table 3.6: CS2 Histotypes

| CodeSet | revHist | n |
|---------|---------------|-----|
| CS2 | CARCINOMA-NOS | 61 |
| CS2 | CCOC | 68 |
| CS2 | CCOC-MCT | 1 |
| CS2 | Cell-Line | 48 |
| CS2 | CTRL | 12 |
| CS2 | ENOC | 30 |
| CS2 | ENOC-CCOC | 7 |
| CS2 | ERROR | 3 |
| CS2 | HGSC | 757 |
| CS2 | HGSC-MCT | 1 |
| CS2 | LGSC | 29 |
| CS2 | MBOT | 20 |
| CS2 | MET-NOP | 21 |
| CS2 | MUC | 61 |
| CS2 | SBOT | 10 |
| CS2 | SQAMOUS | 1 |

Table 3.7: CS3 Histotypes

| CodeSet | revHist | n |
|---------|---|------|
| CS3 | CARCINOMA-NOS | 23 |
| CS3 | Carcinoma, NOS | 2 |
| CS3 | CCOC | 182 |
| CS3 | Cell-Line | 13 |
| CS3 | ENOC | 272 |
| CS3 | HGSC | 2453 |
| CS3 | LGSC | 50 |
| CS3 | MBOT | 3 |
| CS3 | MIXED (ENOC/CCOC) | 1 |
| CS3 | MIXED (ENOC/LGSC) | 1 |
| CS3 | MIXED (HGSC/CCOC) | 1 |
| CS3 | mixed cell | 7 |
| 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 | 3 |
| CS3 | Serous | 2 |
| CS3 | serous LMP | 1 |

Table 3.8: CS1 Training Set Histotypes

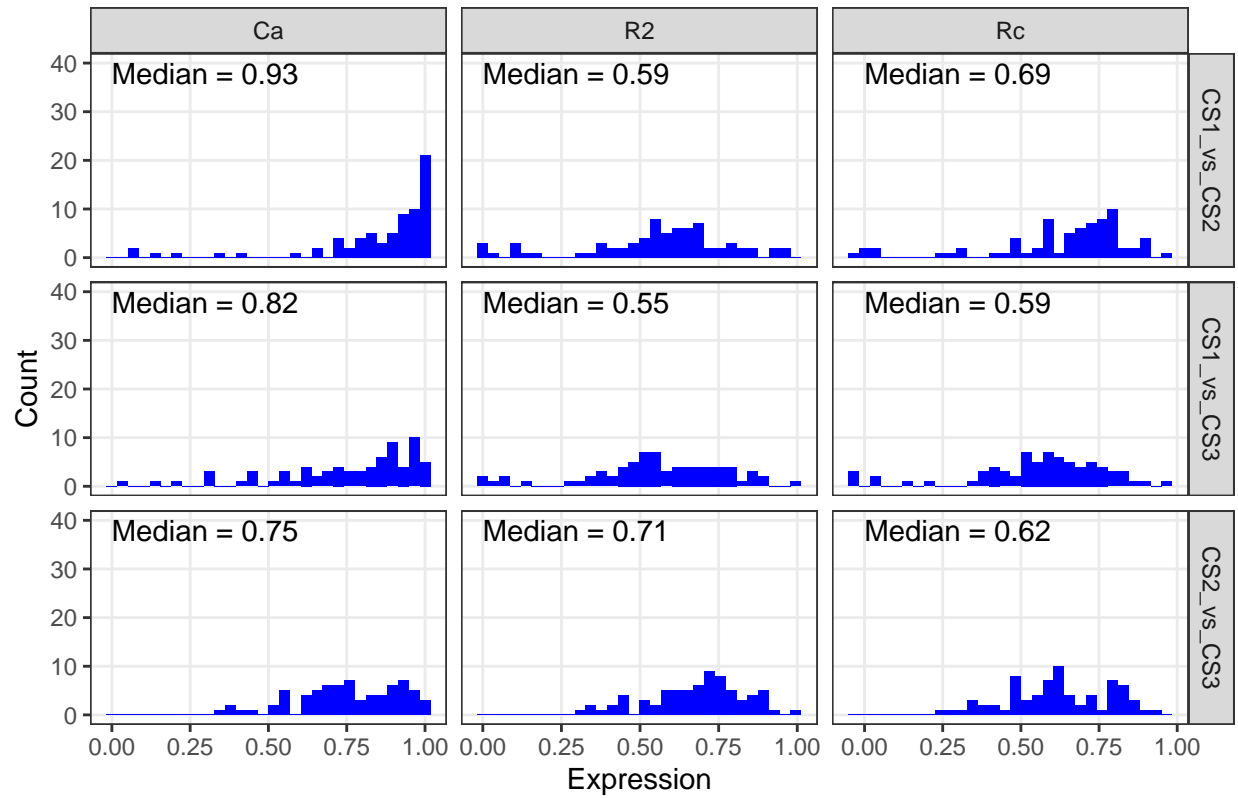
| histotype | n |
|-----------|-----|
| CCC | 57 |
| ENOCa | 59 |
| HGSC | 156 |
| LGSC | 16 |
| MUC | 16 |

Table 3.9: CS2 Training Set Histotypes

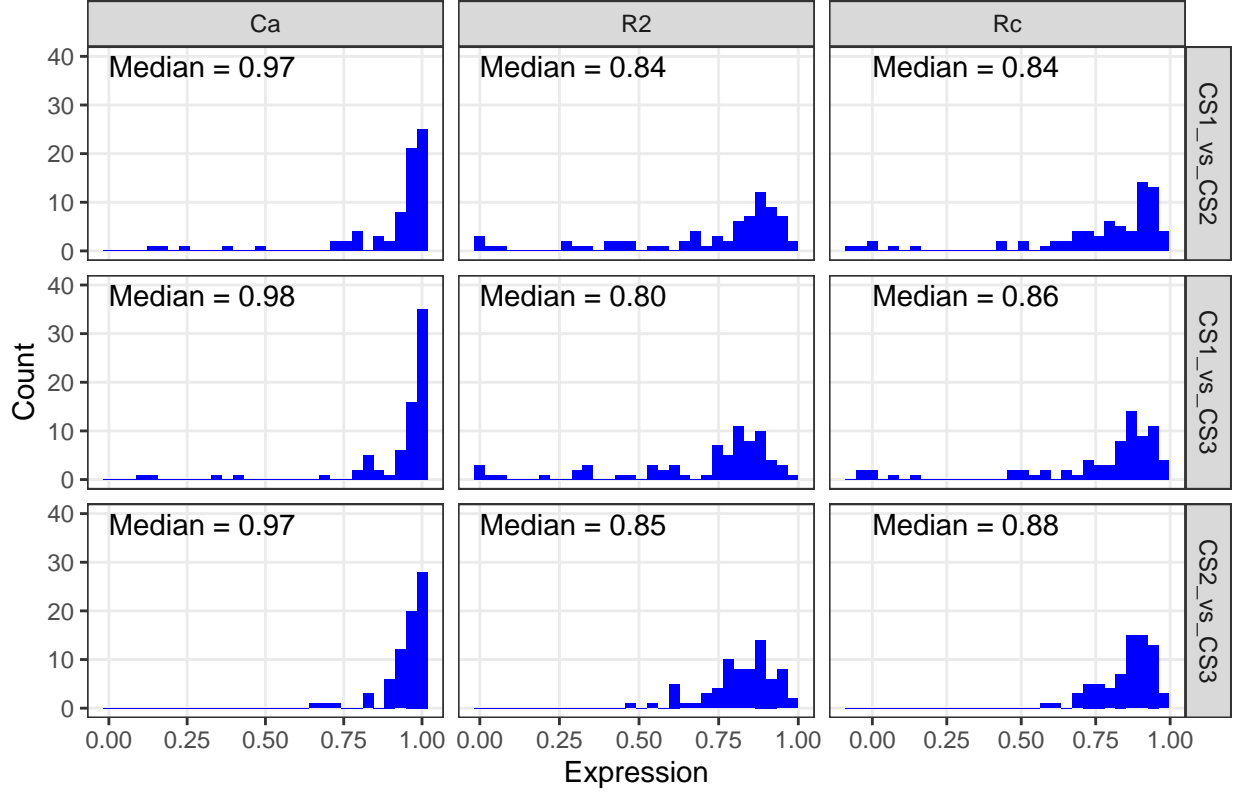
| histotype | n |
|-----------|-----|
| CCOC | 68 |
| ENOC | 30 |
| HGSC | 757 |
| LGSC | 29 |
| MUC | 61 |

3.3 Normalization

Raw Non-Normalized Concordance Measure Distributions



HK genes Normalized Concordance Measure Distributions



3.3.1 Common Samples Method

We employ a new normalization technique using randomly selected samples common to all three CodeSets with a uniform distribution of histotypes as the reference dataset. The number of randomly selected samples ranges from 1-3 per histotype. Hence, the reference dataset has either 5, 10, or 15 samples and we validate on the remaining. Note that ottaID duplicates are collapsed by mean averaging the gene expression. There are n=72 common samples.

CodeSets 1 and 2 are calibrated to CodeSet3 as follows:

- $X^{1(\text{norm})} = X^1 - R^1 + R^3$
- $X^{2(\text{norm})} = X^2 - R^2 + R^3$
- $X^{3(\text{norm})} = X^3$

3.3.1.1 Random3

Randomly choose 3 samples from each of the 5 histotypes as the reference set (n=15). The rest are validated.

Random3 Non-Normalized Concordance Measure Distributions

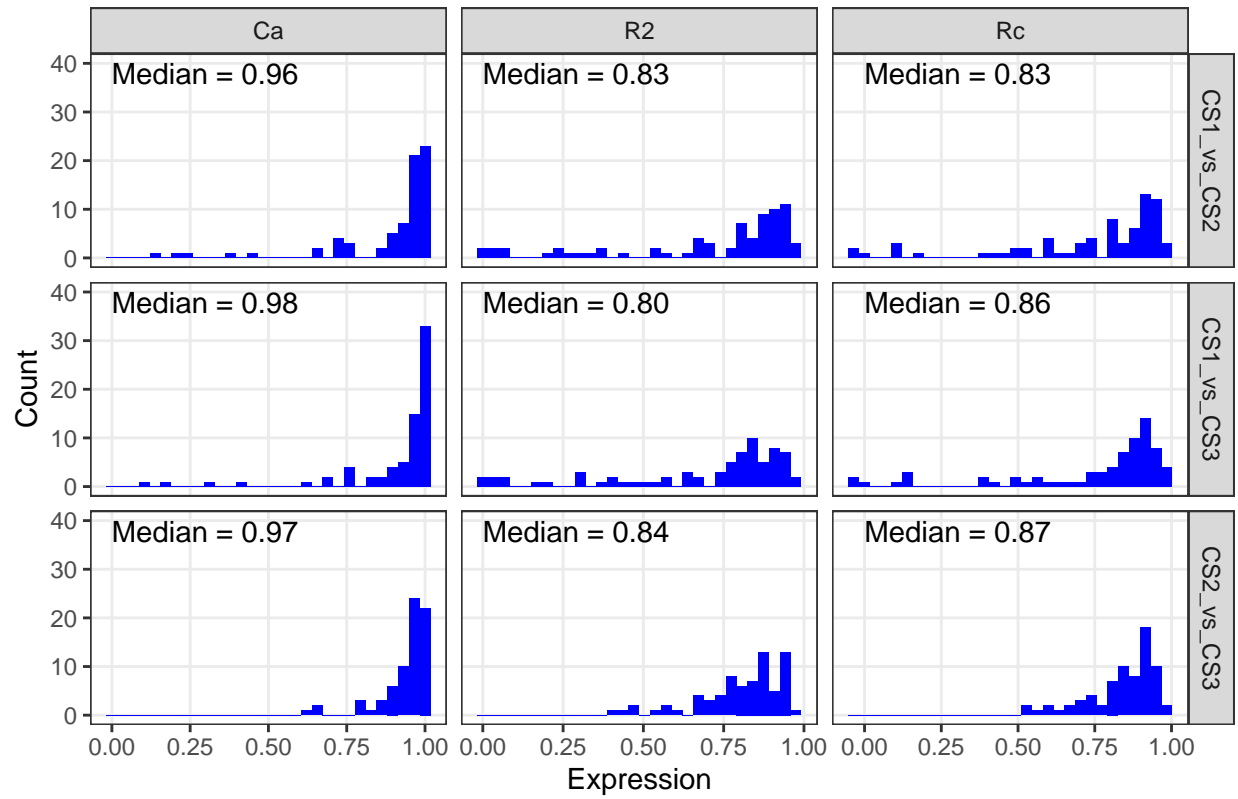


Figure 3.1: Random3 Non-Normalized Concordance Measure Distributions

Random3 Normalized Concordance Measure Distributions

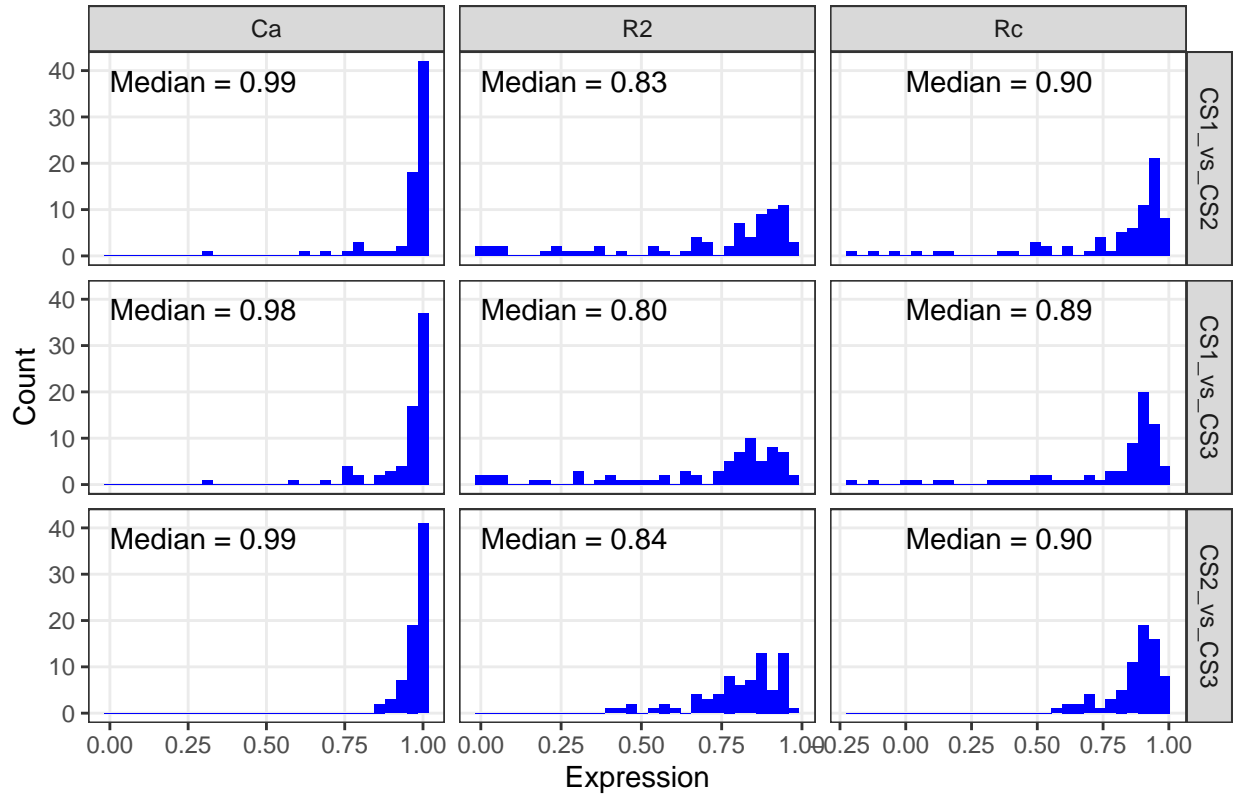


Figure 3.2: Random3 Normalized Concordance Measure Distributions

3.3.1.2 Random2

Randomly choose 2 samples from each of the 5 histotypes as the reference set (n=10). The rest are validated.

Random2 Non-Normalized Concordance Measure Distributions

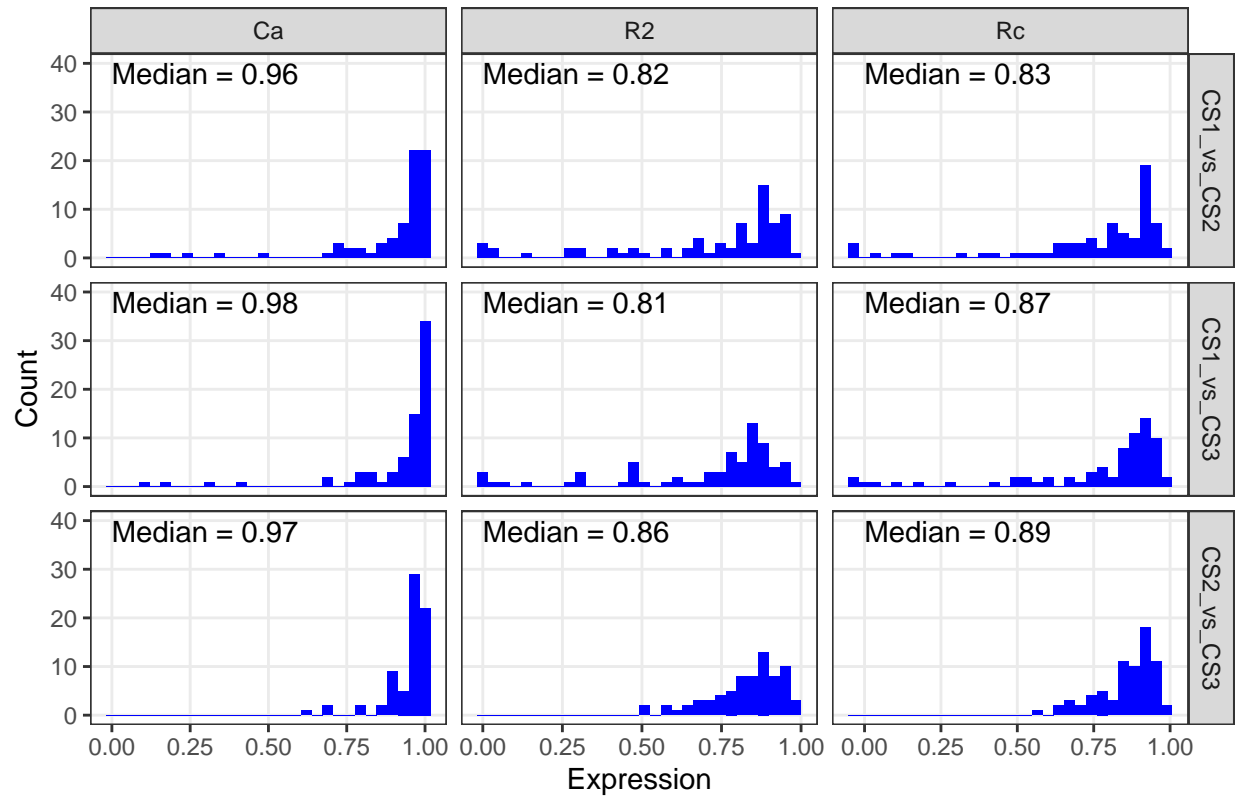


Figure 3.3: Random2 Non-Normalized Concordance Measure Distributions

Random2 Normalized Concordance Measure Distributions

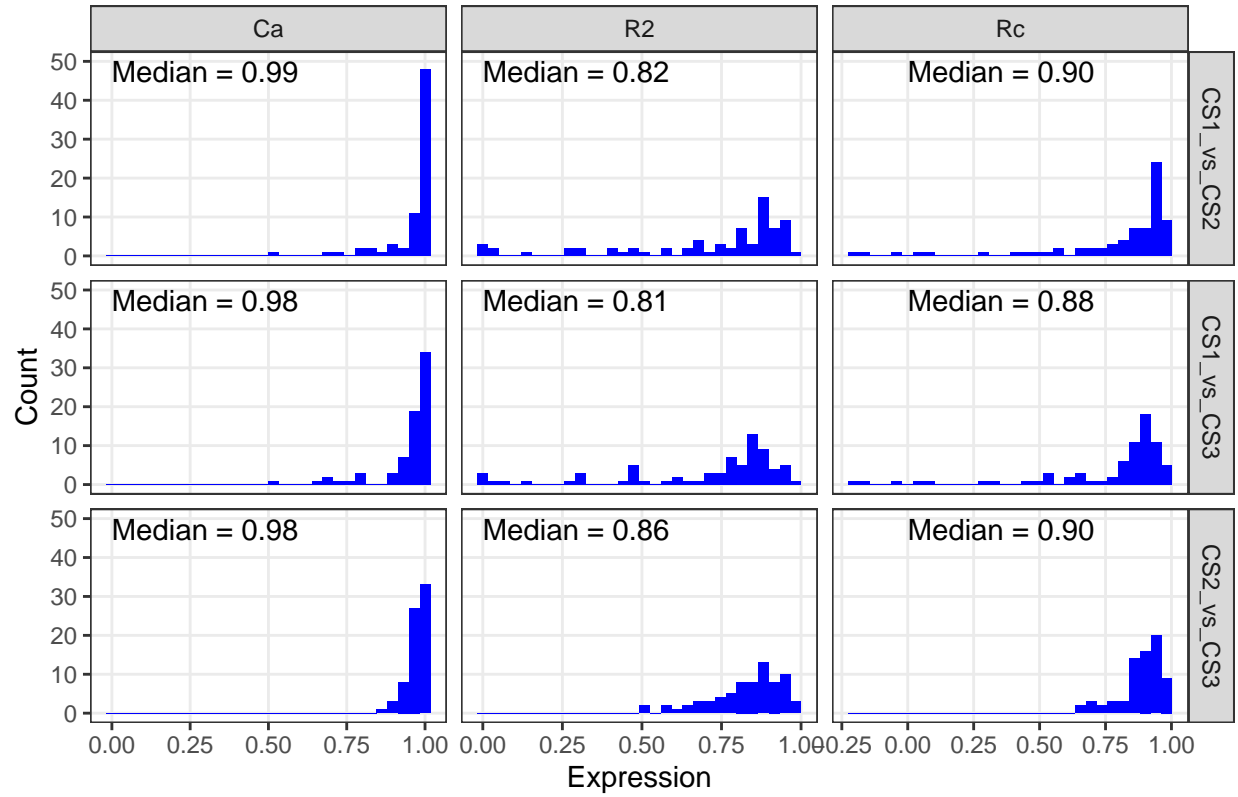


Figure 3.4: Random2 Normalized Concordance Measure Distributions

3.3.1.3 Random1

Randomly choose 1 sample from each of the 5 histotypes as the reference set ($n=5$). The rest are validated.

Random1 Non-Normalized Concordance Measure Distributions

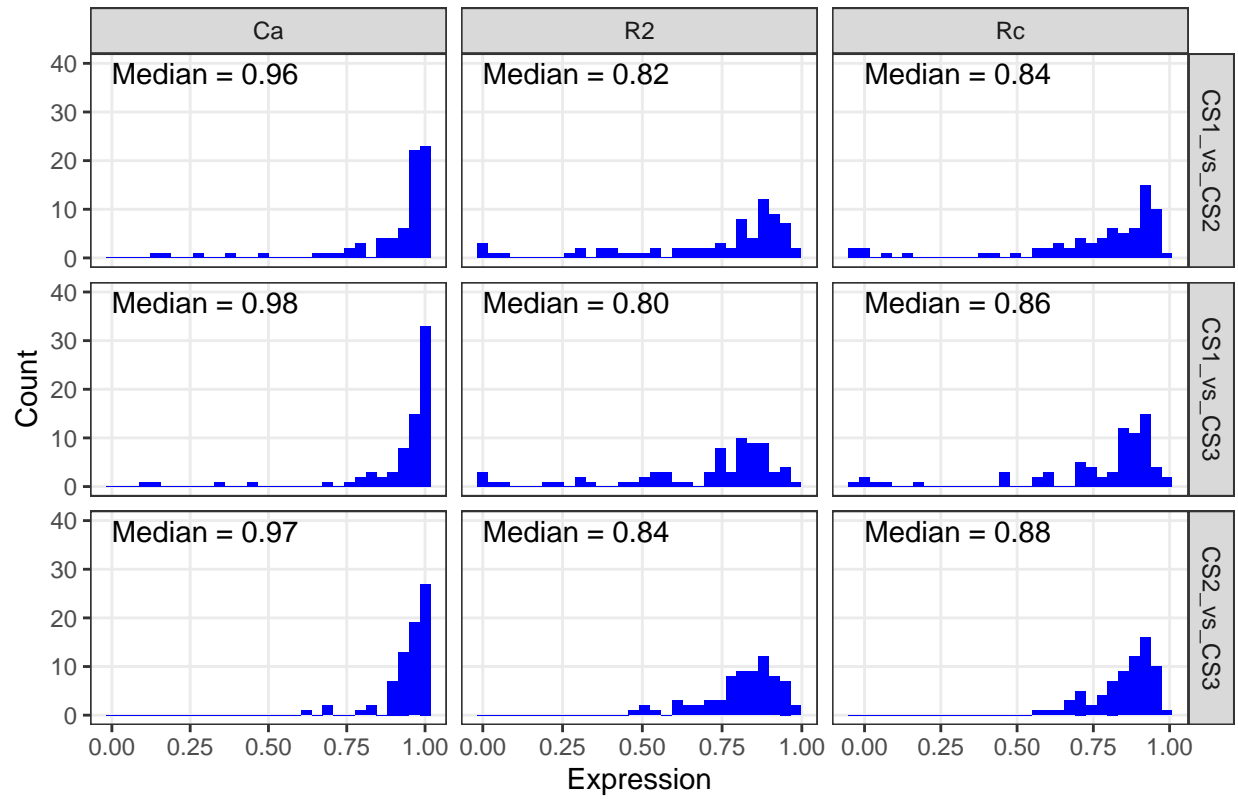


Figure 3.5: Random1 Non-Normalized Concordance Measure Distributions

Table 3.10: Random1 CS1 vs. CS3 Median Concordance Measures by Histotypes

| hist | R2-Non | Ca-Non | Rc-Non | R2-Norm | Ca-Norm | Rc-Norm |
|------|--------|--------|--------|---------|---------|---------|
| CCOC | 1.00 | 0.29 | 0.12 | 1.00 | 0.29 | 0.10 |
| ENOC | 1.00 | 0.54 | 0.54 | 1.00 | 0.62 | 0.62 |
| HGSC | 0.79 | 0.98 | 0.85 | 0.79 | 0.97 | 0.87 |
| LGSC | 0.96 | 0.89 | 0.82 | 0.96 | 0.91 | 0.87 |
| MUC | 0.77 | 0.86 | 0.68 | 0.77 | 0.81 | 0.63 |

Table 3.11: Random1 CS2 vs. CS3 Median Concordance Measures by Histotypes

| hist | R2-Non | Ca-Non | Rc-Non | R2-Norm | Ca-Norm | Rc-Norm |
|------|--------|--------|--------|---------|---------|---------|
| CCOC | 1.00 | 0.23 | 0.08 | 1.00 | 0.27 | 0.16 |
| ENOC | 1.00 | 0.63 | 0.61 | 1.00 | 0.61 | 0.57 |
| HGSC | 0.83 | 0.96 | 0.86 | 0.83 | 0.98 | 0.89 |
| LGSC | 0.98 | 0.92 | 0.90 | 0.98 | 0.95 | 0.93 |
| MUC | 0.68 | 0.77 | 0.55 | 0.68 | 0.86 | 0.61 |

Random1 Normalized Concordance Measure Distributions

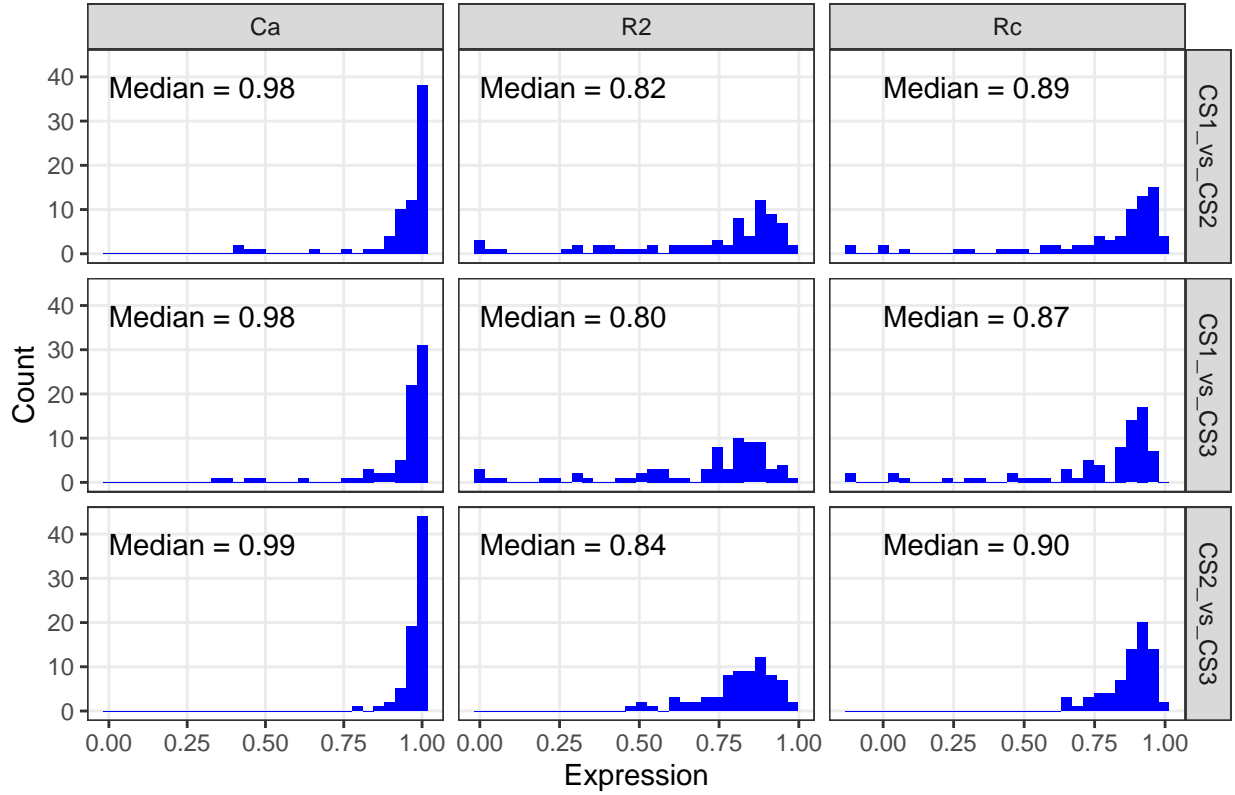


Figure 3.6: Random1 Normalized Concordance Measure Distributions

In Tables 3.10 and 3.11, we calculate the concordance measures for CS1 vs. CS3 and CS2 vs. CS3, respectively. The measures are calculated for both non-normalized and normalized datasets (CS1, CS2), and split by histotype.

Table 3.12: CS2 vs CS3 Random1 Normalized 100 Runs of Summary Concordance Measures

| Metric | Min | Median | Max | SD |
|--------|-------|--------|-------|-------|
| Ca | 0.956 | 0.981 | 0.989 | 0.007 |
| R2 | 0.830 | 0.847 | 0.870 | 0.007 |
| Rc | 0.867 | 0.891 | 0.907 | 0.008 |

Because Random1 is a random selection of reference samples, we want to assess the variability of the concordance measures by repeating Random1 on different selections and observing the distribution of the medians.

CS2 vs. CS3 Random1 Normalized 100 Runs of Median Concordance Measures

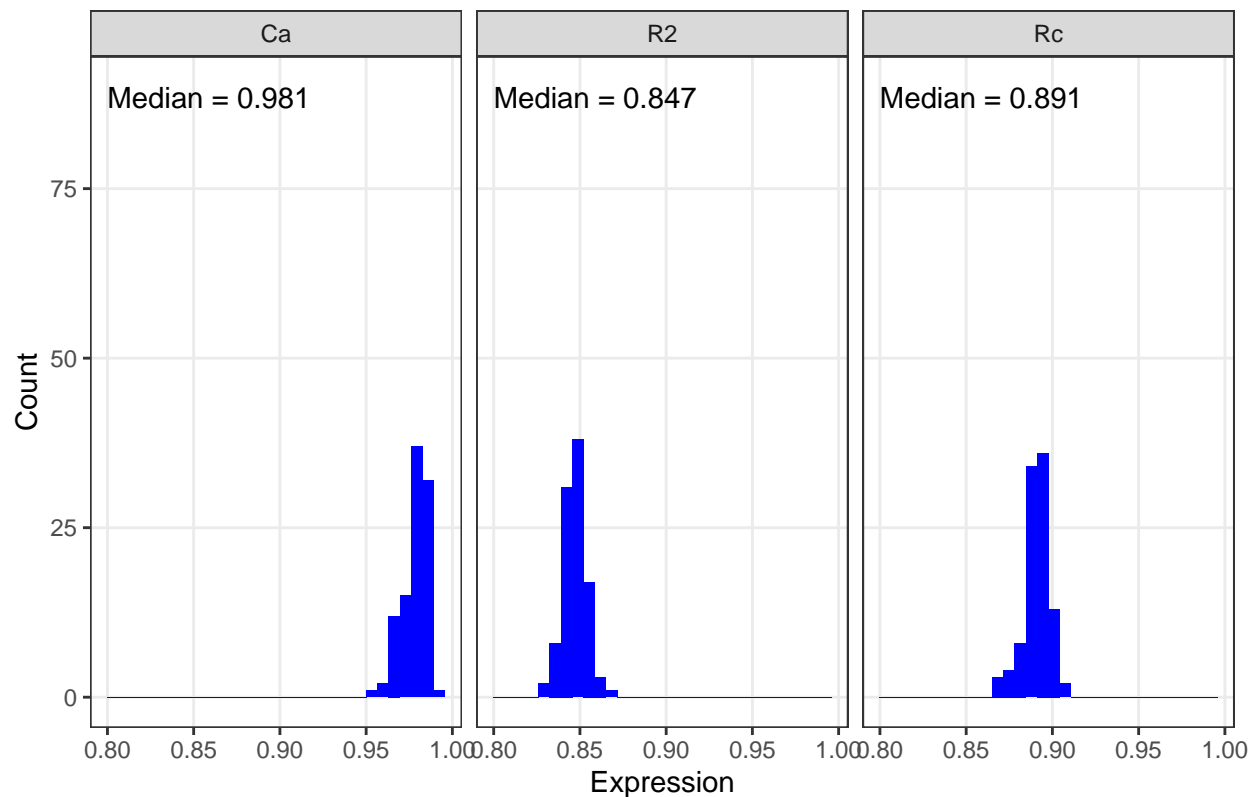


Figure 3.7: CS2 vs. CS3 Random1 Normalized 100 Runs of Median Concordance Measures

3.3.1.4 Random3 HGSC

Randomly choose $n=3$ HGSC samples as the reference set, and use the rest as validation. This was tried in lieu of the fact that some non-HGSC histotypes have at most $n=3$ samples in total, so using Random3 or even Random2 would leave no samples remaining in the validation set for these histotypes.

Table 3.13: Random3 HGSC CS1 vs. CS3 Median Concordance Measures by Histotypes

| hist | R2-Non | Ca-Non | Rc-Non | R2-Norm | Ca-Norm | Rc-Norm |
|------|--------|--------|--------|---------|---------|---------|
| CCOC | 0.62 | 0.62 | 0.32 | 0.62 | 0.68 | 0.27 |
| ENOC | 0.88 | 0.76 | 0.66 | 0.88 | 0.77 | 0.70 |
| HGSC | 0.77 | 0.97 | 0.85 | 0.77 | 0.99 | 0.87 |
| LGSC | 0.94 | 0.85 | 0.80 | 0.94 | 0.90 | 0.84 |
| MUC | 0.74 | 0.92 | 0.72 | 0.74 | 0.93 | 0.78 |

Table 3.14: Random3 HGSC CS2 vs. CS3 Median Concordance Measures by Histotypes

| hist | R2-Non | Ca-Non | Rc-Non | R2-Norm | Ca-Norm | Rc-Norm |
|------|--------|--------|--------|---------|---------|---------|
| CCOC | 0.66 | 0.56 | 0.35 | 0.66 | 0.59 | 0.42 |
| ENOC | 0.85 | 0.76 | 0.66 | 0.85 | 0.85 | 0.76 |
| HGSC | 0.82 | 0.96 | 0.86 | 0.82 | 0.99 | 0.90 |
| LGSC | 0.97 | 0.95 | 0.92 | 0.97 | 0.92 | 0.90 |
| MUC | 0.74 | 0.89 | 0.72 | 0.74 | 0.93 | 0.72 |

Random3 HGSC Normalized Concordance Measure Distributions

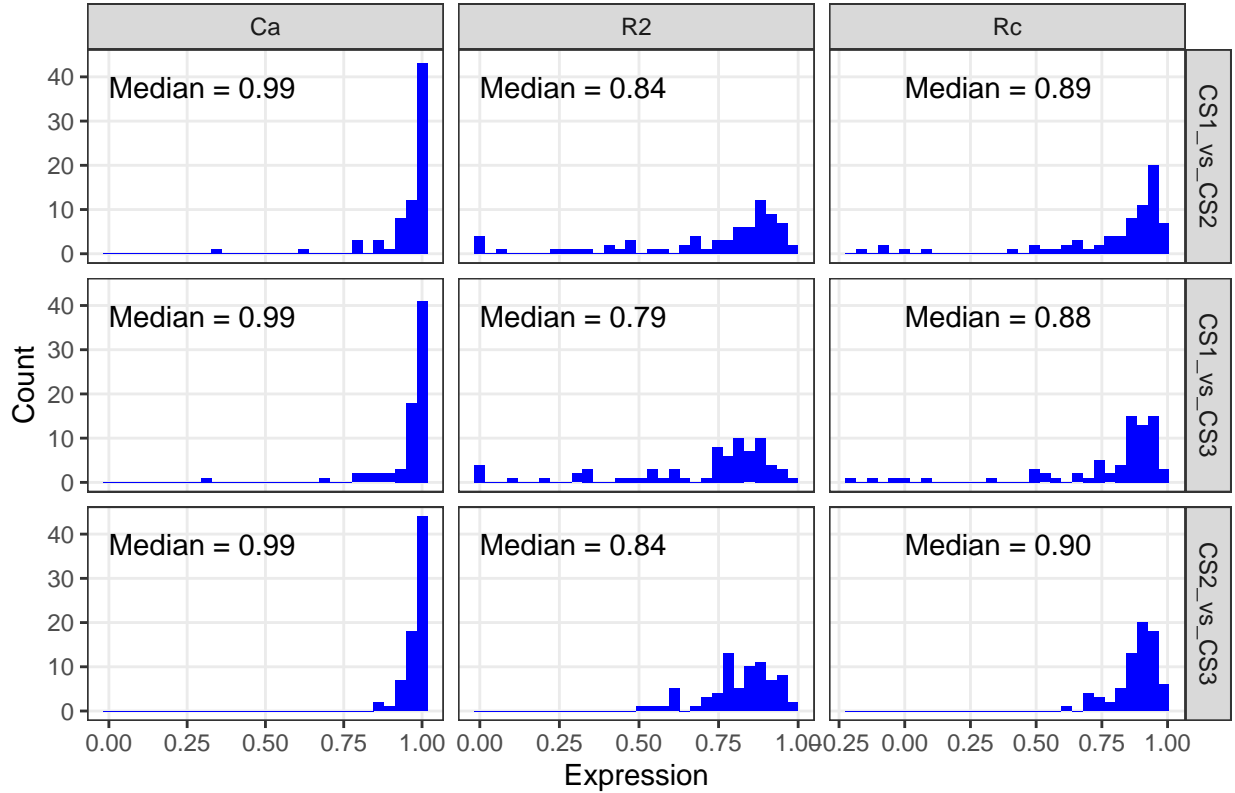


Figure 3.8: Random3 HGSC Normalized Concordance Measure Distributions

In Tables 3.13 and 3.14, we calculate the concordance measures for CS1 vs. CS3 and CS2 vs. CS3, respectively. The measures are calculated for both non-normalized and normalized datasets (CS1, CS2), and split by histotype.

3.3.1.5 Random1 for Sites

We use the Random1 method to normalize CS3-USC and CS3-AOC to CS3-VAN. There aren't enough samples in the USC and AOC cohorts to perform Random2 or Random3.

Cross-Site Random1 Non-Normalized Concordance Measure Distribution:

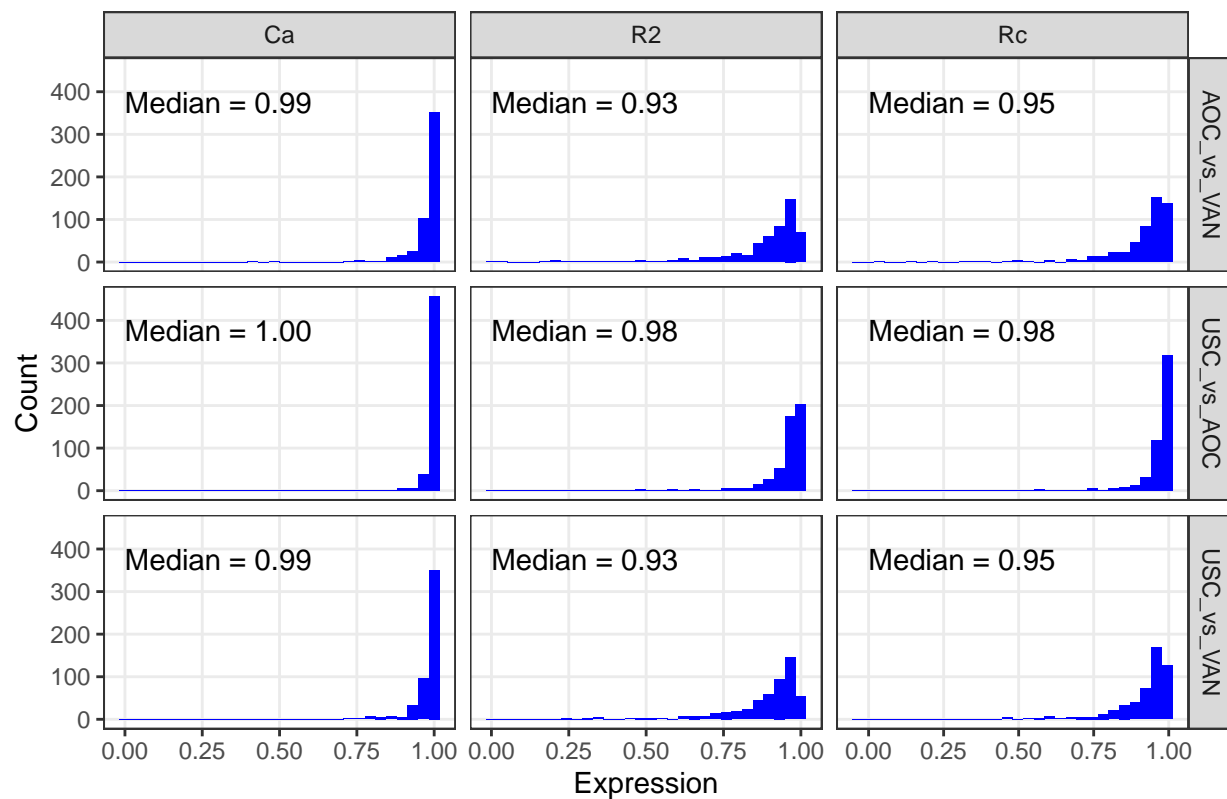
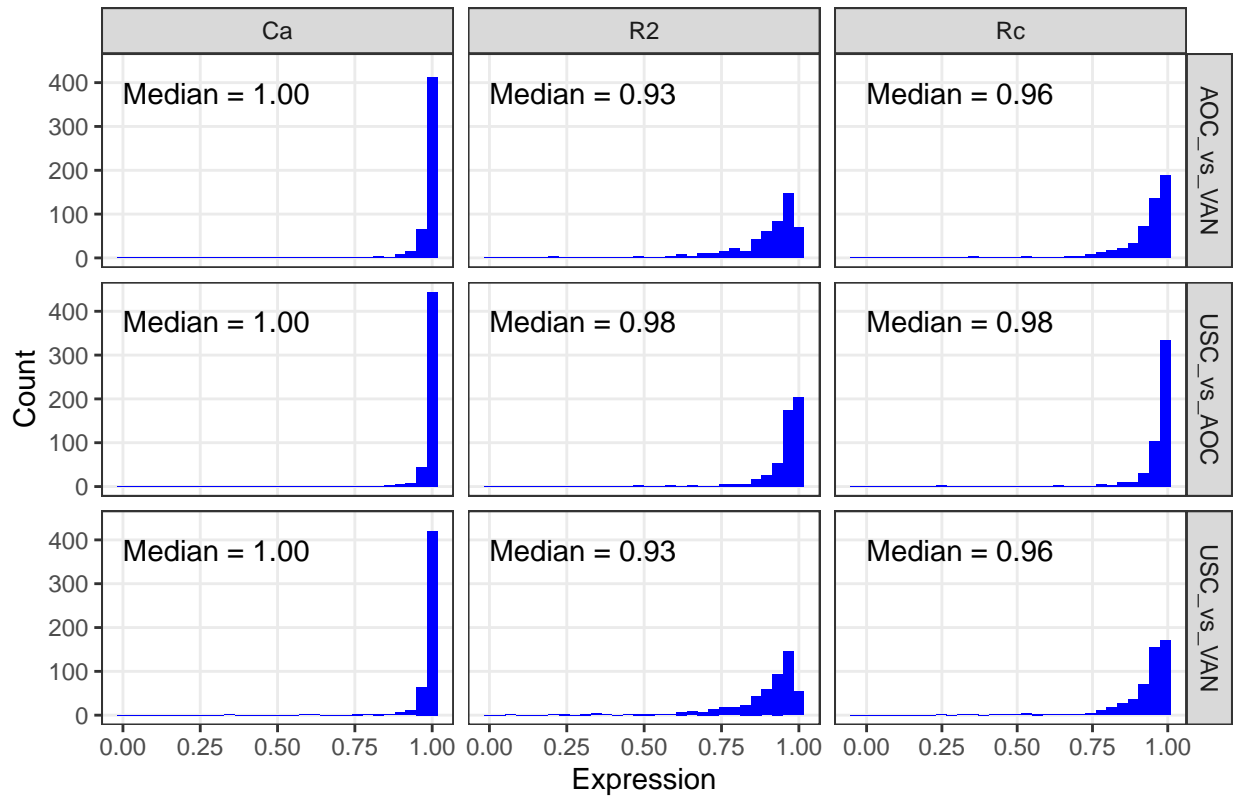


Figure 3.9: Cross-Site Random1 Non-Normalized Concordance Measure Distributions

Cross-Site Random1 Normalized Concordance Measure Distributions



3.3.2 Pools Method

3.3.2.1 CS2 vs. CS3

CodeSet2 contains 12 ref pool samples (Pool 1 = 4, Pool 2 = 4, Pool 3 = 4). CodeSet3 contains 22 ref pool samples (Pool 1 = 12, Pool 2 = 5, Pool 3 = 5). n=84 common samples.

CodeSet2 is calibrated to CodeSet3 as follows:

$$X^2(\text{norm}) = X^2 - R^2 + R^3$$

$$X^3(\text{norm}) = X^3$$

CS2Non vs. CS2Pools Concordance Measure Distributions

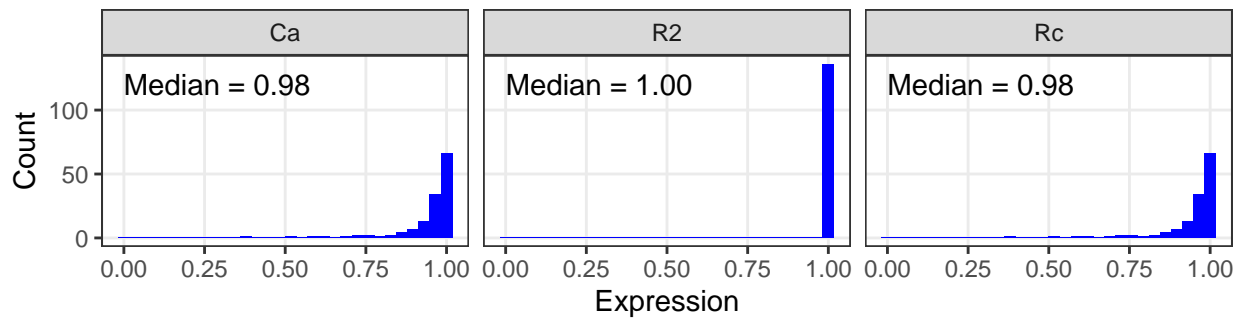


Figure 3.10: CS2Non vs. CS2Pools Concordance Measure Distributions

Table 3.15: Pools Non-Normalized CS2 vs. CS3 Median Concordance Measures by Histotypes

| hist | R2 | Ca | Rc |
|------|------|------|------|
| CCOC | 0.66 | 0.53 | 0.26 |
| ENOC | 0.88 | 0.74 | 0.63 |
| HGSC | 0.77 | 0.94 | 0.80 |
| LGSC | 0.98 | 0.95 | 0.92 |
| MUC | 0.74 | 0.86 | 0.68 |

Table 3.16: Pools Normalized CS2 vs. CS3 Median Concordance Measures by Histotypes

| hist | R2 | Ca | Rc |
|------|------|------|------|
| CCOC | 0.66 | 0.60 | 0.32 |
| ENOC | 0.88 | 0.76 | 0.68 |
| HGSC | 0.77 | 0.94 | 0.81 |
| LGSC | 0.98 | 0.95 | 0.93 |
| MUC | 0.74 | 0.91 | 0.71 |

CS2 Non-Normalized Pools vs. CS3 Concordance Measure Distributions

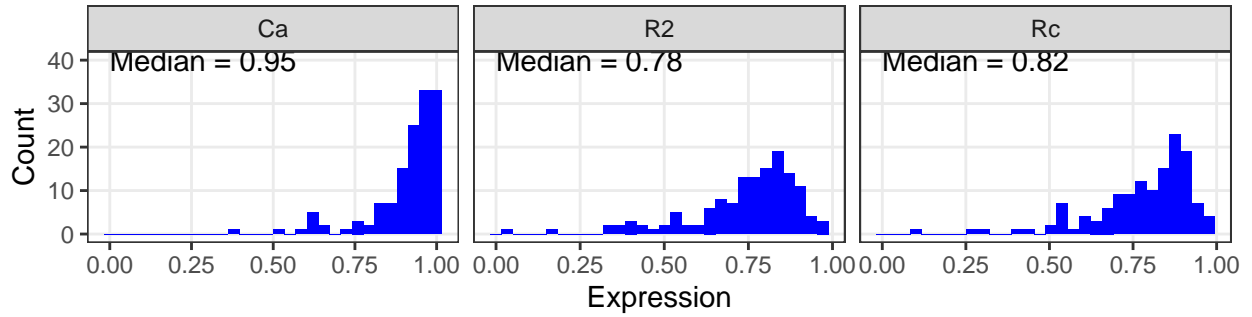


Figure 3.11: CS2 Non-Normalized Pools vs. CS3 Concordance Measure Distributions

CS2 Normalized Pools vs. CS3 Concordance Measure Distributions

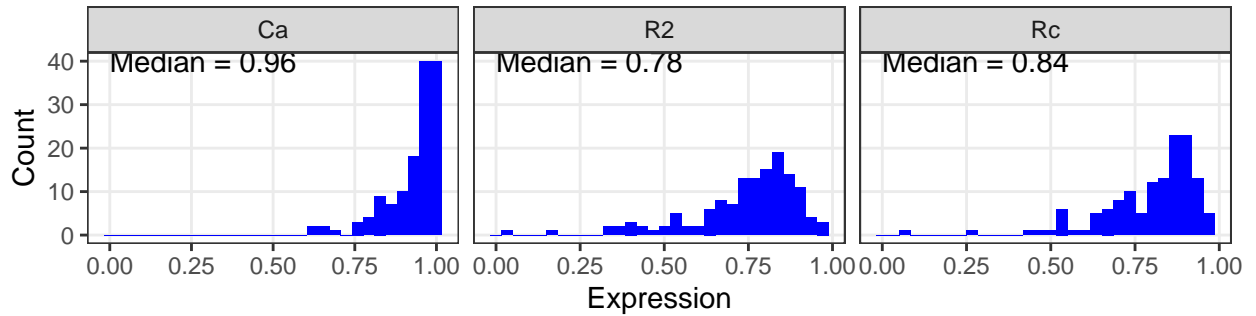


Figure 3.12: CS2 Normalized Pools vs. CS3 Concordance Measure Distributions

3.3.2.2 USC vs. VAN

In CodeSet 3, we normalize the USC and AOC cohorts to the VAN cohort which is used as the reference dataset.

USC–Non vs. USC–Pools Concordance Measure Distributions

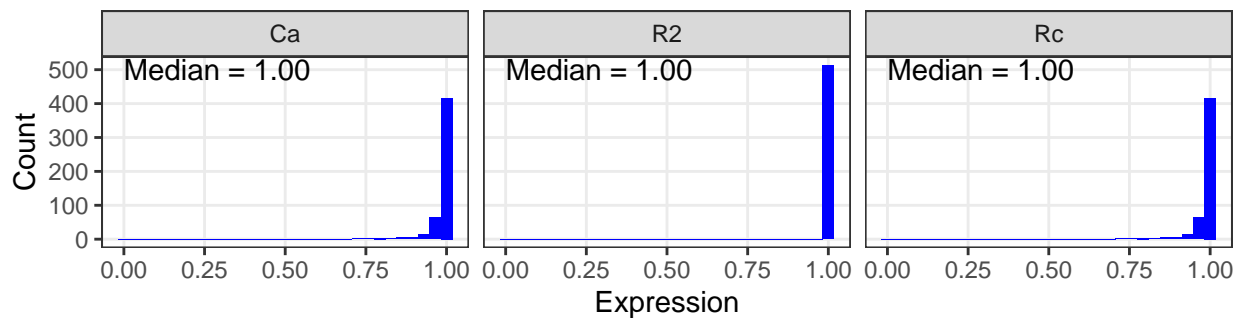


Figure 3.13: USC–Non vs. USC–Pools Concordance Measure Distributions

USC–Non vs. VAN–Non Concordance Measure Distributions

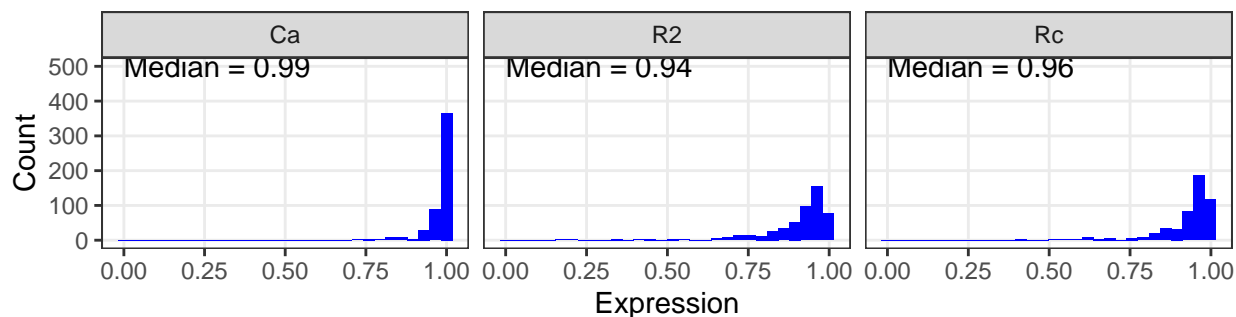


Figure 3.14: USC–Non vs. VAN–Non Concordance Measure Distributions

USC–Pools vs. VAN–Non Concordance Measure Distributions

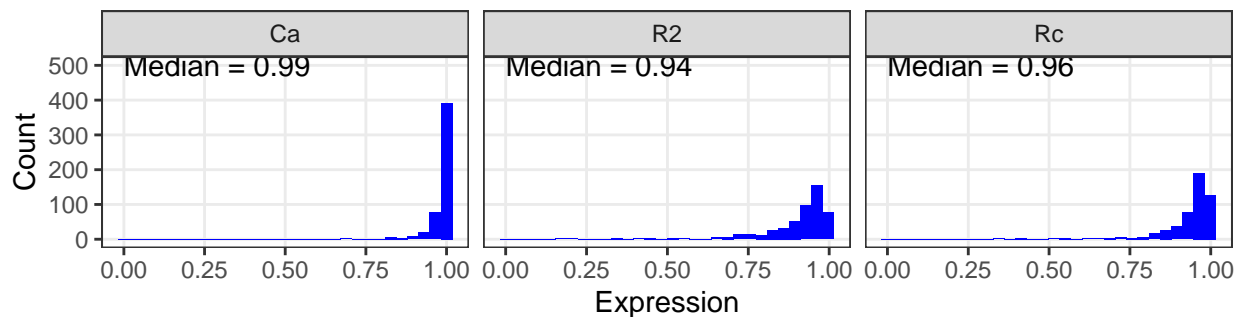


Figure 3.15: USC–Pools vs. VAN–Non Concordance Measure Distributions

USC vs. VAN Comparisons of Concordance Measure Distributions

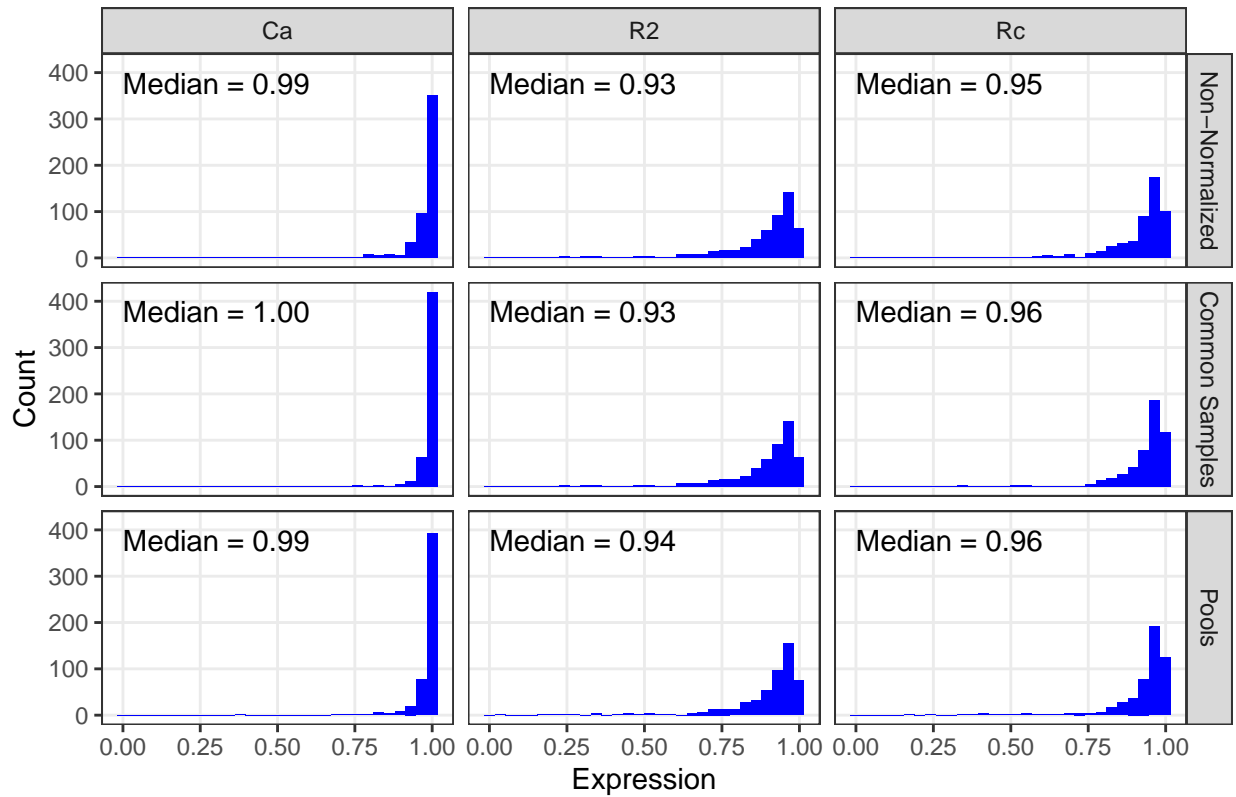


Figure 3.16: USC vs. VAN Comparisons of Concordance Measure Distributions

3.3.2.3 AOC vs. VAN

AOC-Non vs. AOC-Pools Concordance Measure Distributions

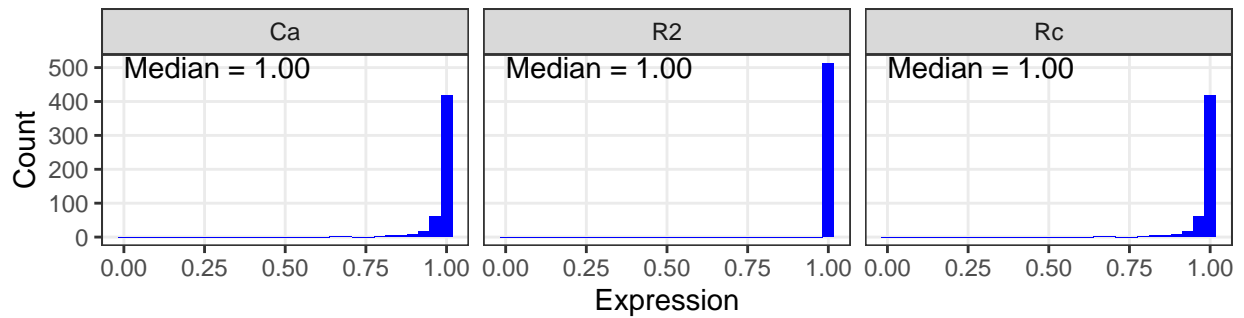


Figure 3.17: AOC-Non vs. AOC-Pools Concordance Measure Distributions

AOC–Non vs. VAN–Non Concordance Measure Distributions

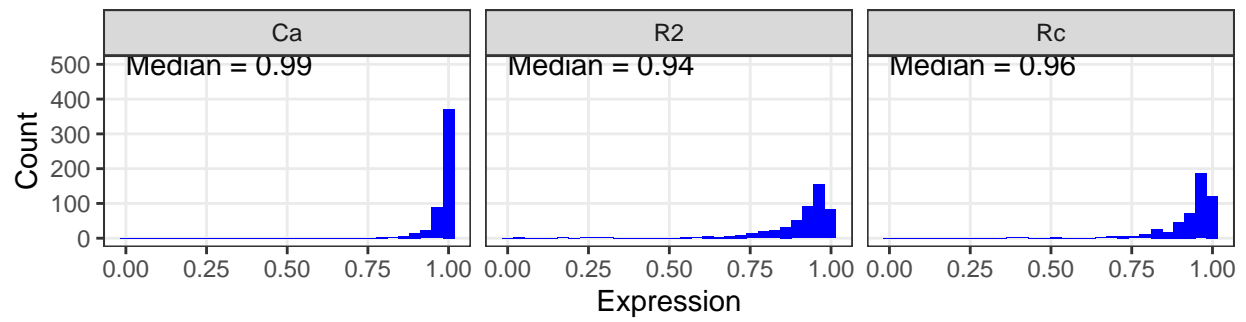


Figure 3.18: AOC–Non vs. VAN–Non Concordance Measure Distributions

AOC–Pools vs. VAN–Non Concordance Measure Distributions

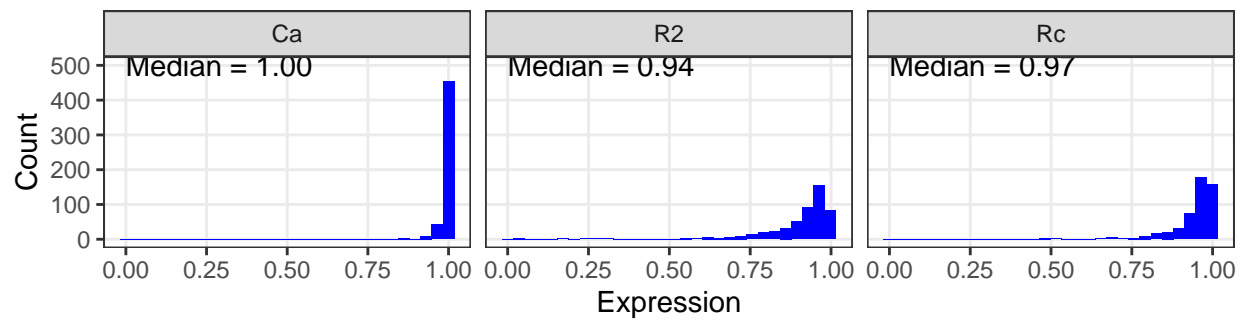


Figure 3.19: AOC–Pools vs. VAN–Non Concordance Measure Distributions

AOC vs. VAN Comparisons of Concordance Measure Distributions

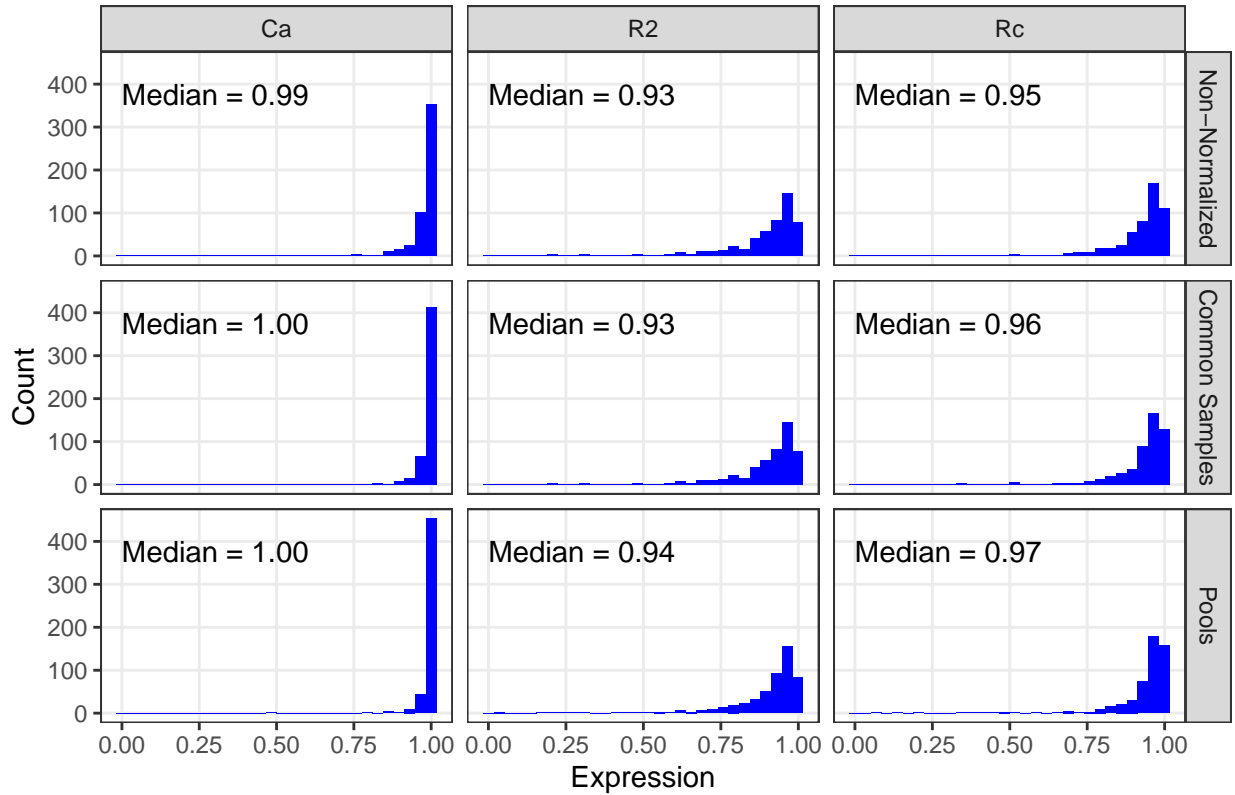


Figure 3.20: AOC vs. VAN Comparisons of Concordance Measure Distributions

3.3.3 Common Samples vs. Pools Comparison

Since only CS2 and CS3 have pools, we make three comparisons between these two CodeSets:

- Non-Normalized
- Common Samples Method
- Pools Method

Table 3.17: Random3 Samples Comparisons Statistics by Histotypes

| hist | R2-Non | Ca-Non | Rc-Non | R2-Common | Ca-Common | Rc-Common | R2-Pools | Ca-Pools | Rc-Pools |
|------|--------|--------|--------|-----------|-----------|-----------|----------|----------|----------|
| HGSC | 0.84 | 0.96 | 0.86 | 0.84 | 0.99 | 0.90 | 0.84 | 0.96 | 0.86 |
| LGSC | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| MUC | 1.00 | 0.49 | 0.44 | 1.00 | 0.62 | 0.52 | 1.00 | 0.46 | 0.42 |

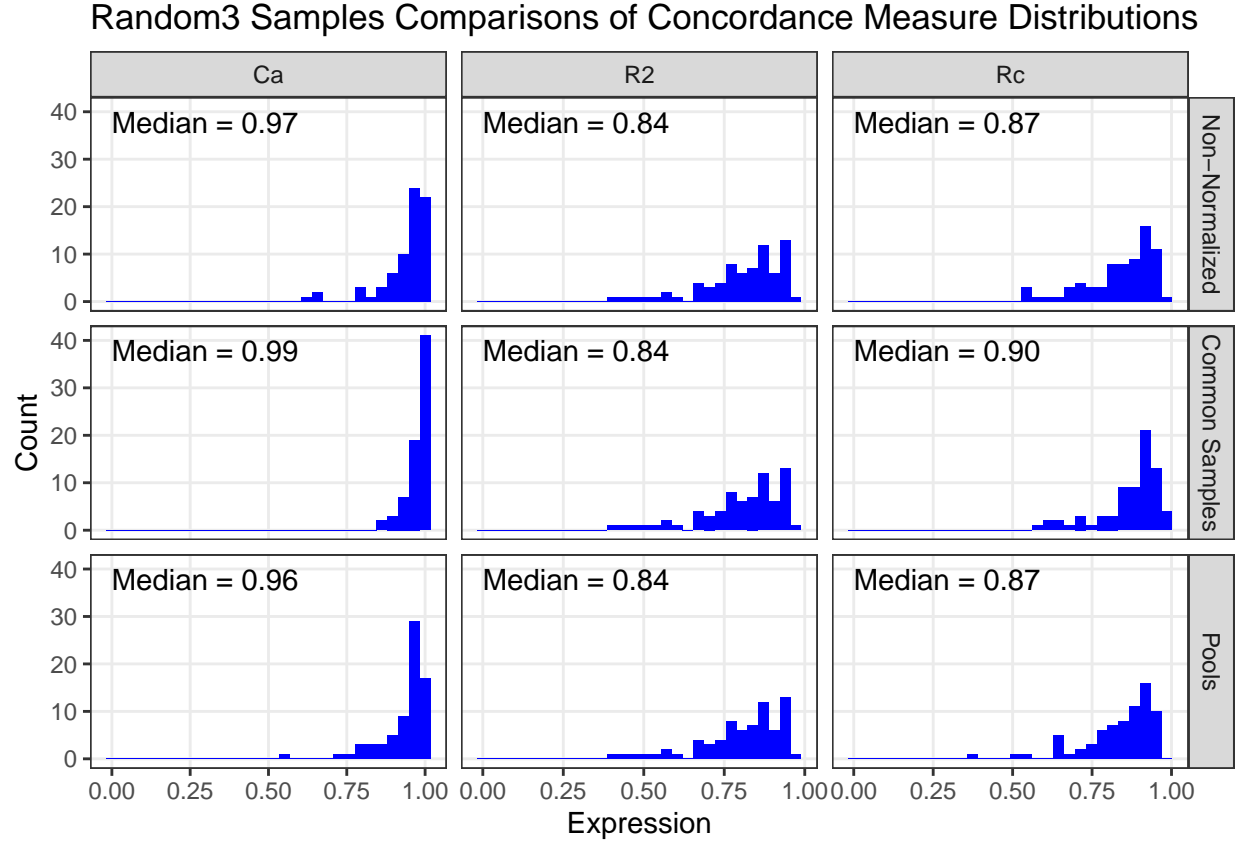


Figure 3.21: Random3 Samples Comparisons of Concordance Measure Distributions

Table 3.18: Random2 Samples Comparisons Statistics by Histotypes

| hist | R2-Non | Ca-Non | Rc-Non | R2-Common | Ca-Common | Rc-Common | R2-Pools | Ca-Pools | Rc-Pools |
|------|--------|--------|--------|-----------|-----------|-----------|----------|----------|----------|
| CCOC | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| ENOC | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| HGSC | 0.84 | 0.96 | 0.87 | 0.84 | 0.98 | 0.89 | 0.84 | 0.96 | 0.86 |
| LGSC | 1.00 | 0.88 | 0.87 | 1.00 | 0.88 | 0.88 | 1.00 | 0.85 | 0.85 |
| MUC | 0.97 | 0.95 | 0.91 | 0.97 | 0.94 | 0.90 | 0.97 | 0.96 | 0.92 |

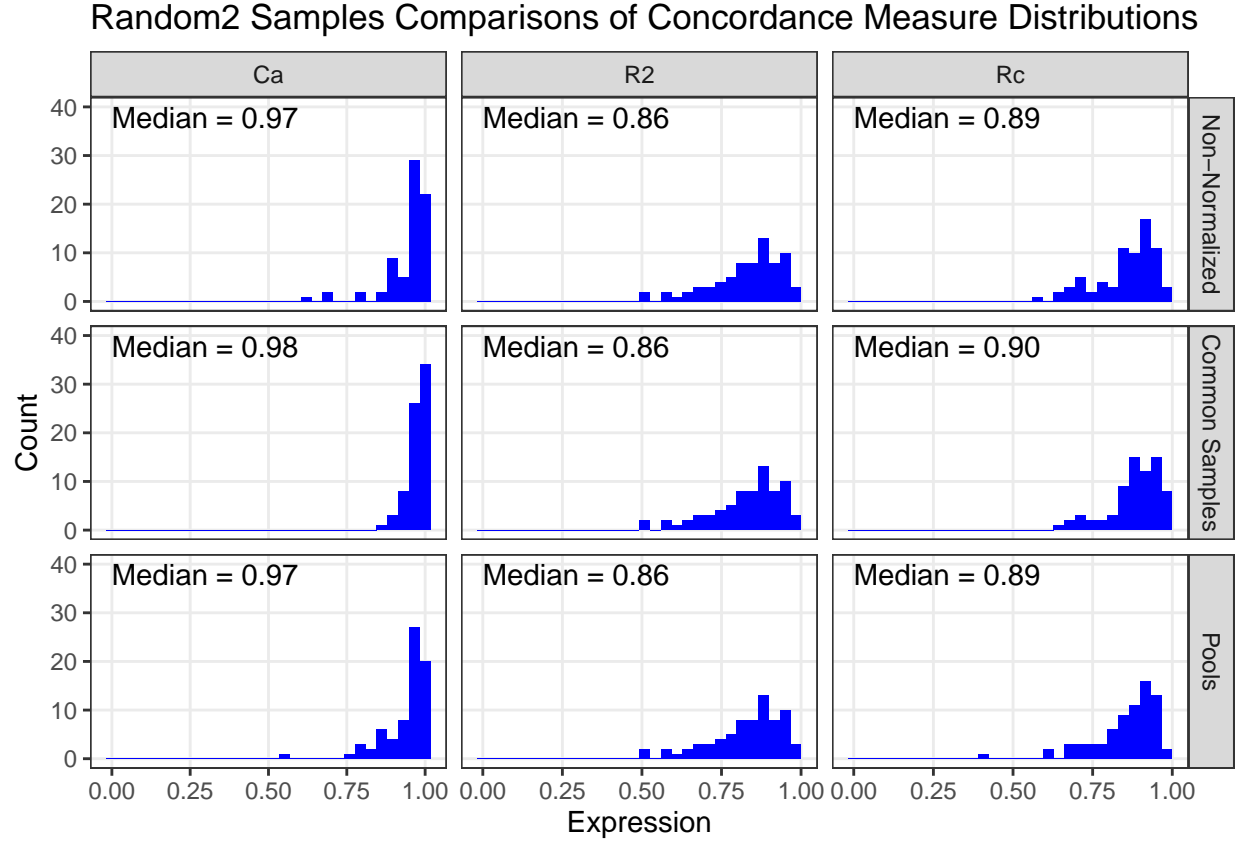


Figure 3.22: Random2 Samples Comparisons of Concordance Measure Distributions

Table 3.19: Random1 Samples Comparisons Statistics by Histotypes

| hist | R2-Non | Ca-Non | Rc-Non | R2-Common | Ca-Common | Rc-Common | R2-Pools | Ca-Pools | Rc-Pools |
|------|--------|--------|--------|-----------|-----------|-----------|----------|----------|----------|
| CCOC | 1.00 | 0.23 | 0.08 | 1.00 | 0.27 | 0.16 | 1.00 | 0.15 | 0.09 |
| ENOC | 1.00 | 0.63 | 0.61 | 1.00 | 0.61 | 0.57 | 1.00 | 0.61 | 0.61 |
| HGSC | 0.83 | 0.96 | 0.86 | 0.83 | 0.98 | 0.89 | 0.83 | 0.96 | 0.86 |
| LGSC | 0.98 | 0.92 | 0.90 | 0.98 | 0.95 | 0.93 | 0.98 | 0.92 | 0.90 |
| MUC | 0.68 | 0.77 | 0.55 | 0.68 | 0.86 | 0.61 | 0.68 | 0.78 | 0.51 |

Table 3.20: DSC for CS2 vs CS3 Comparisons

| Comparison | DSC | pval |
|--------------------|-------|-------|
| CS2-None vs CS3 | 0.126 | 0.032 |
| CS2-Random1 vs CS3 | 0.084 | 0.475 |
| CS2-Pools vs CS3 | 0.159 | 0.005 |

Random1 Samples Comparisons of Concordance Measure Distributions

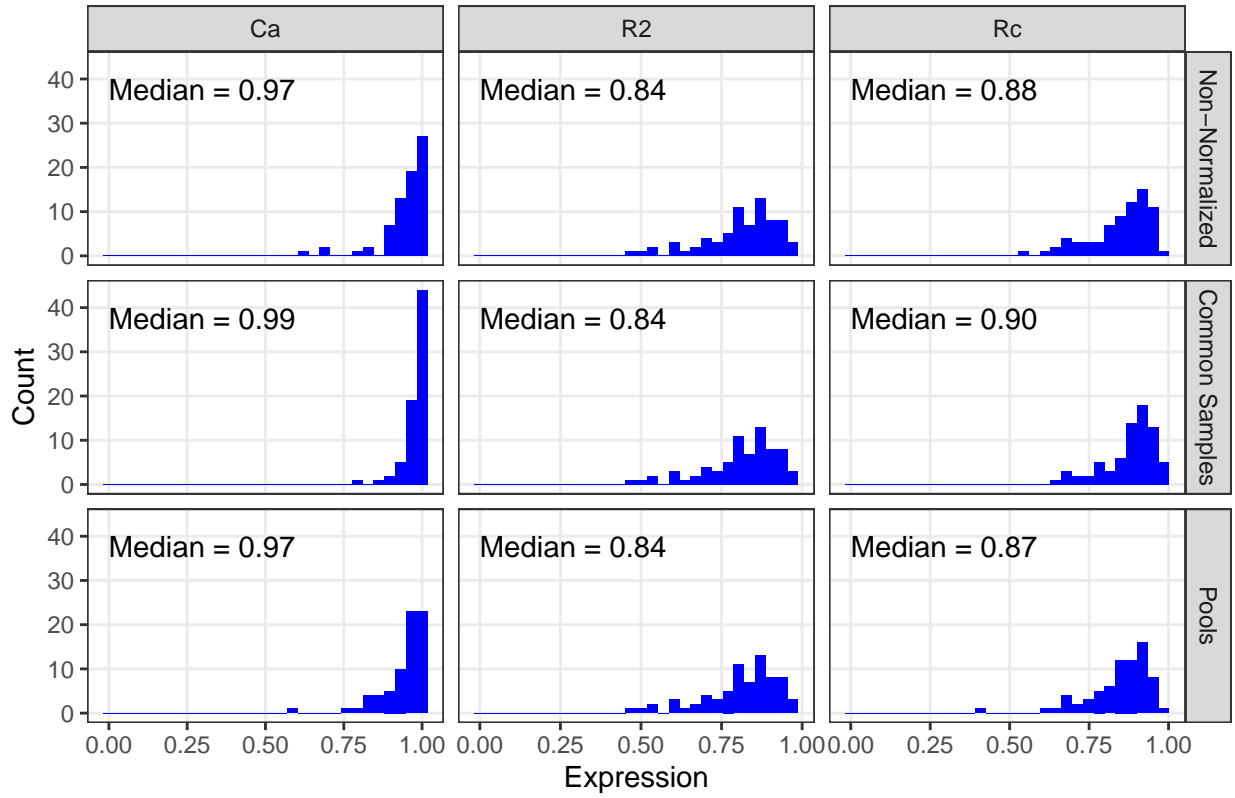


Figure 3.23: Random1 Samples Comparisons of Concordance Measure Distributions

3.3.4 CodeSet Chaining

3.3.4.1 CS1, CS2, CS3

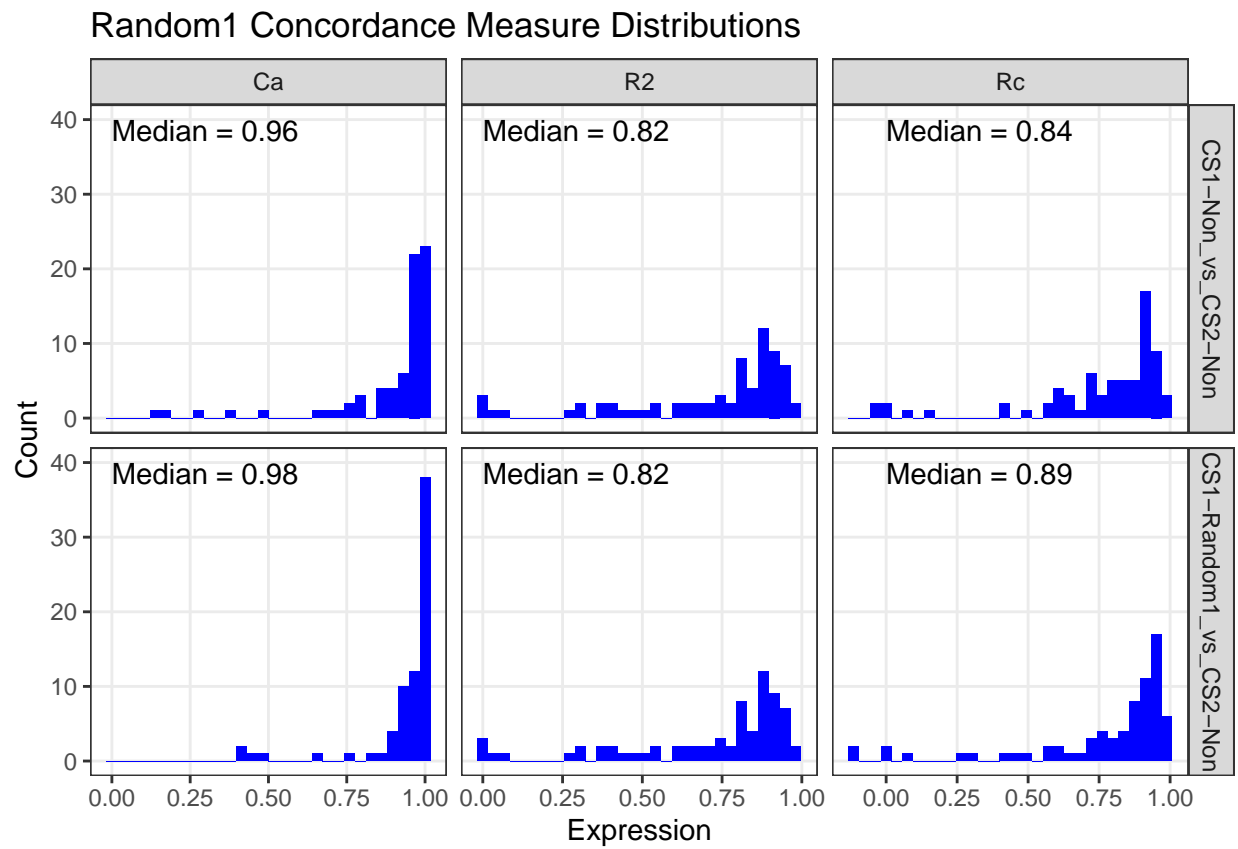


Figure 3.24: Random1 Concordance Measure Distributions

Random1 + Pools Concordance Measure Distributions

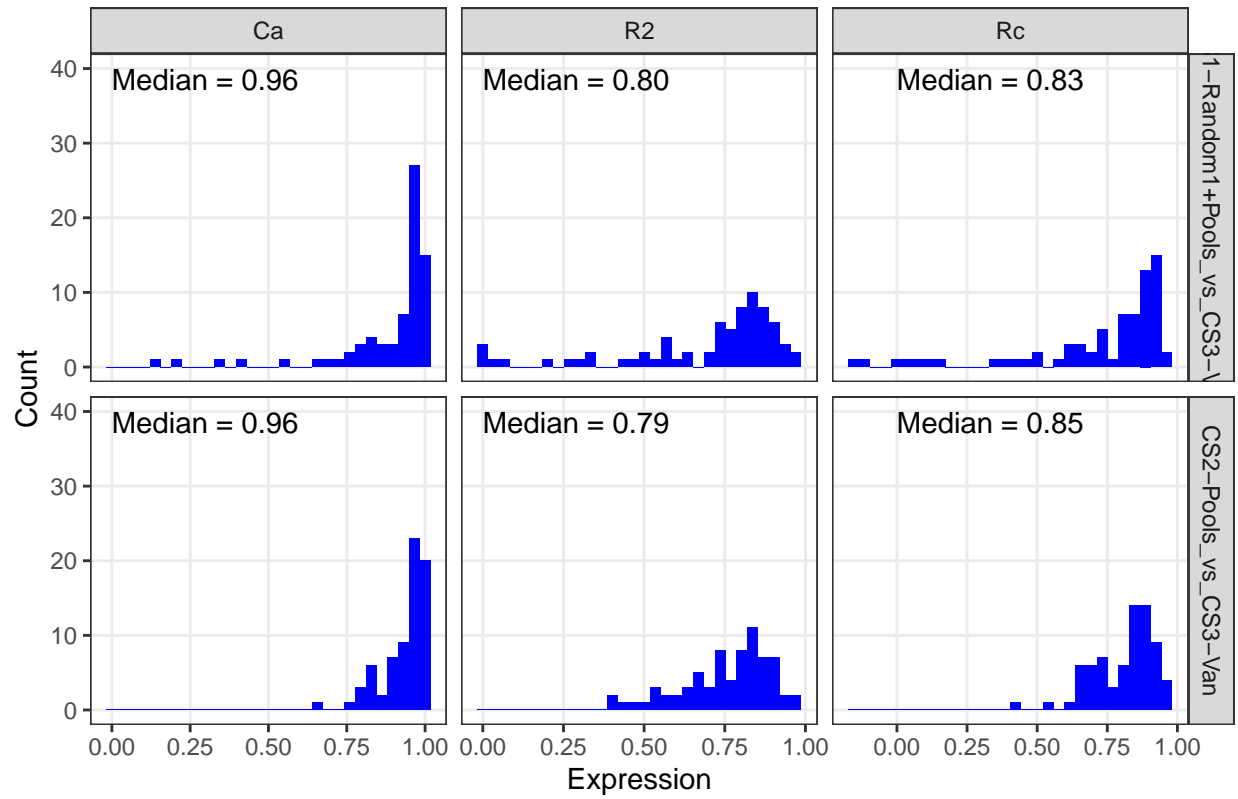


Figure 3.25: Random1 + Pools Concordance Measure Distributions

CS1 CodeSet Chaining Concordance Measure Distributions

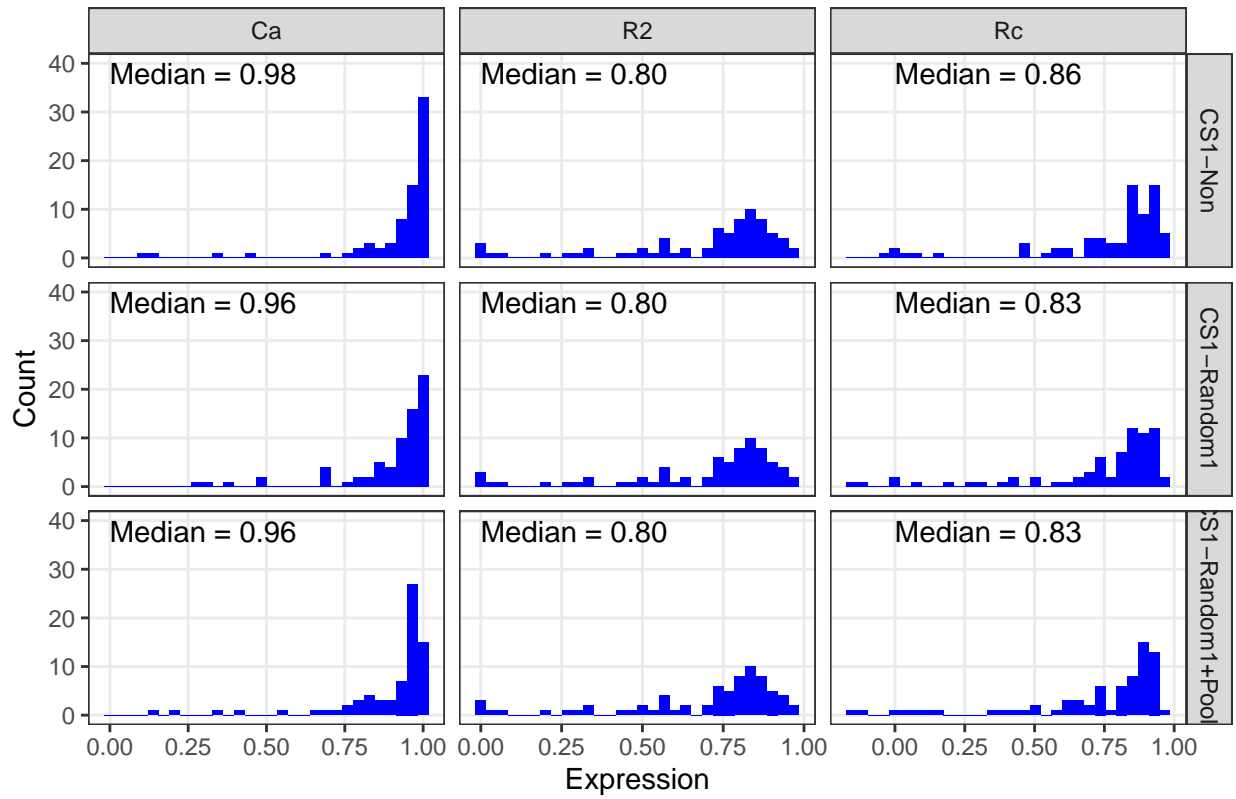


Figure 3.26: CS1 CodeSet Chaining Concordance Measure Distributions

Table 3.21: DSC for CS1 vs CS3 Comparisons

| Comparison | DSC | pval |
|--------------------------|-------|------|
| CS1-None vs CS3 | 0.265 | 0 |
| CS1-Random1+Pools vs CS3 | 0.351 | 0 |
| CS1-Random1 vs CS3 | 0.229 | 0 |

CS1 CodeSet Chaining Concordance Measure Distributions 2

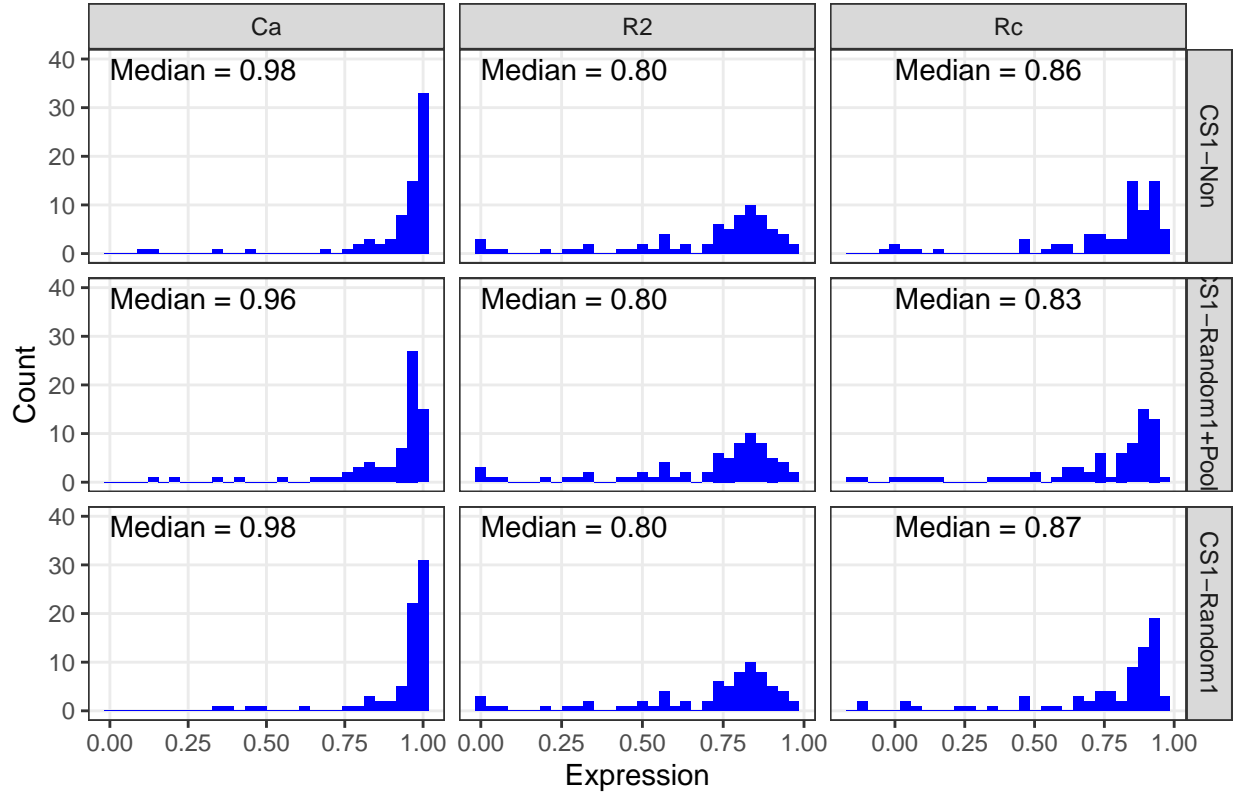


Figure 3.27: CS1 CodeSet Chaining Concordance Measure Distributions 2

Pairwise Genes by Top/Bottom 3 Rc for CS1 vs. CS3

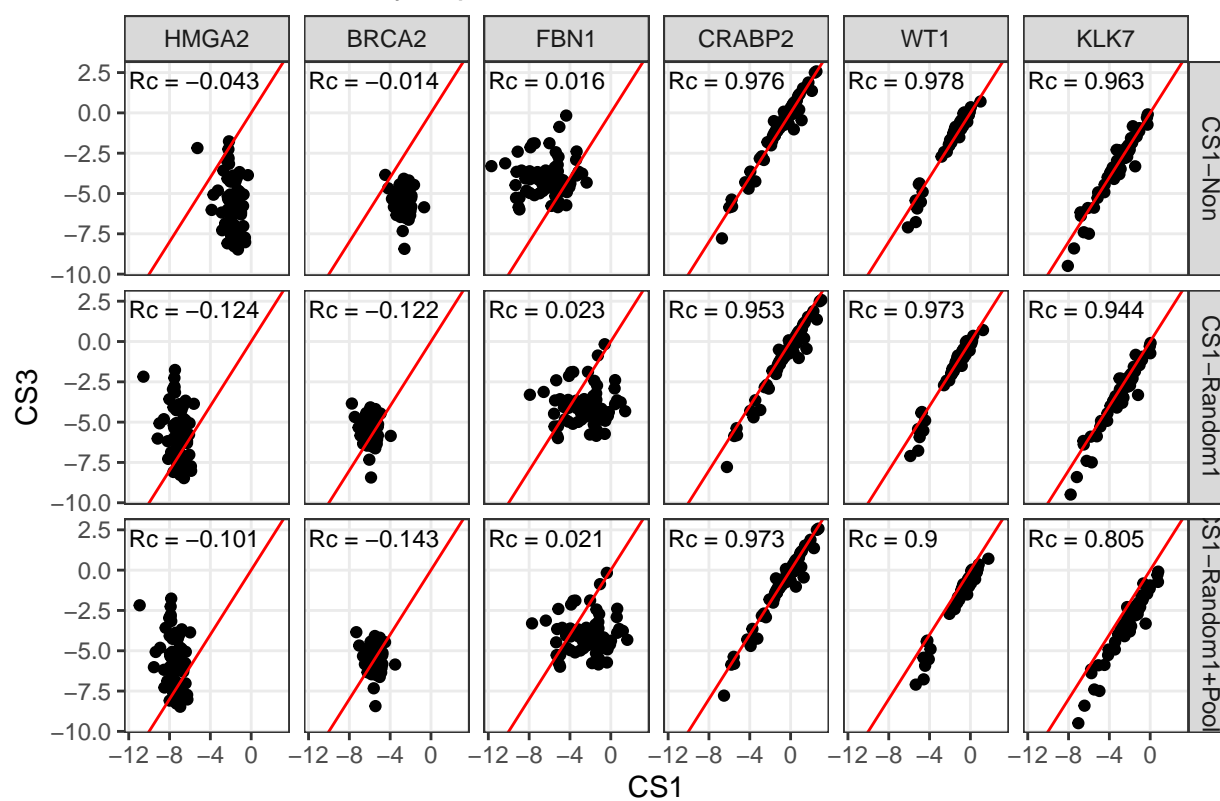


Figure 3.28: Pairwise Genes by Top/Bottom 3 Rc for CS1 vs. CS3

CS2 CodeSet Chaining Concordance Measure Distributions

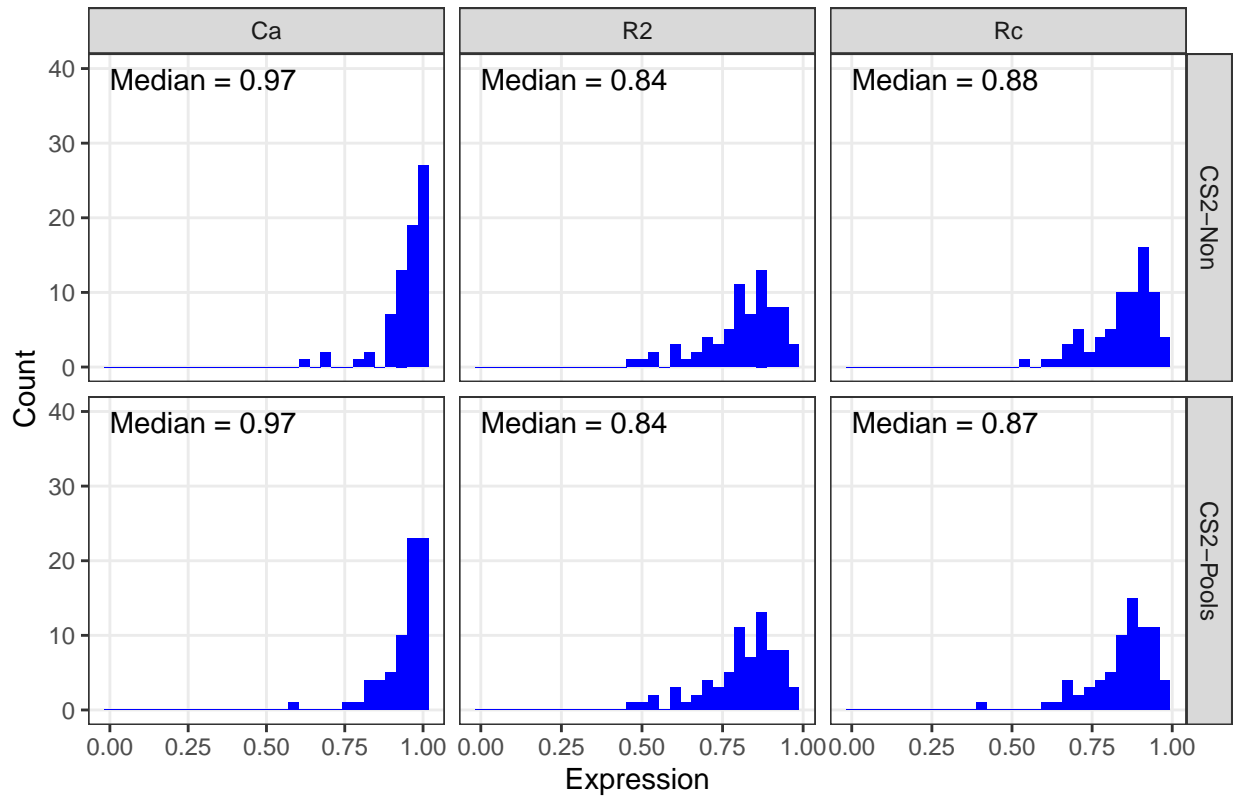


Figure 3.29: CS2 CodeSet Chaining Concordance Measure Distributions

3.3.4.2 CS3, CS4, CS5 using Set B/A

CS5 Set B/A Chaining Concordance Measure Distributions

Samples=72, Genes=55

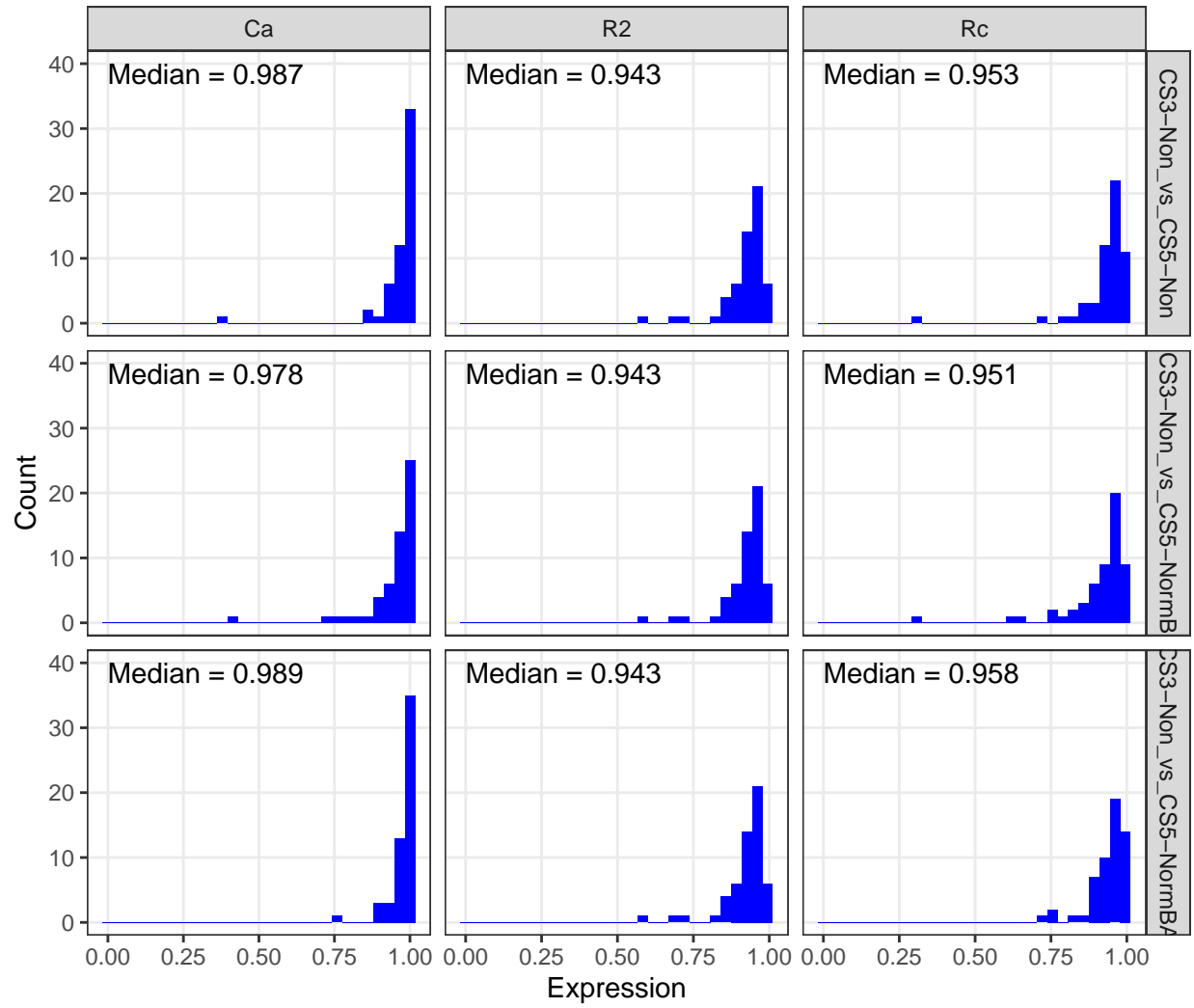


Figure 3.30: CS5 Set B/A Chaining Concordance Measure Distributions

Table 3.22: DSC for CS3 vs CS5 Set B/A Comparisons

| Comparisons | DSC | pval |
|-------------------|-------|-------|
| CS3 vs CS5-None | 0.126 | 0.101 |
| CS3 vs CS5-NormBA | 0.087 | 0.320 |
| CS3 vs CS5-NormA | 0.144 | 0.047 |

CS5 Set B/A Chaining Concordance Measure Distributions 2

Samples=72, Genes=55

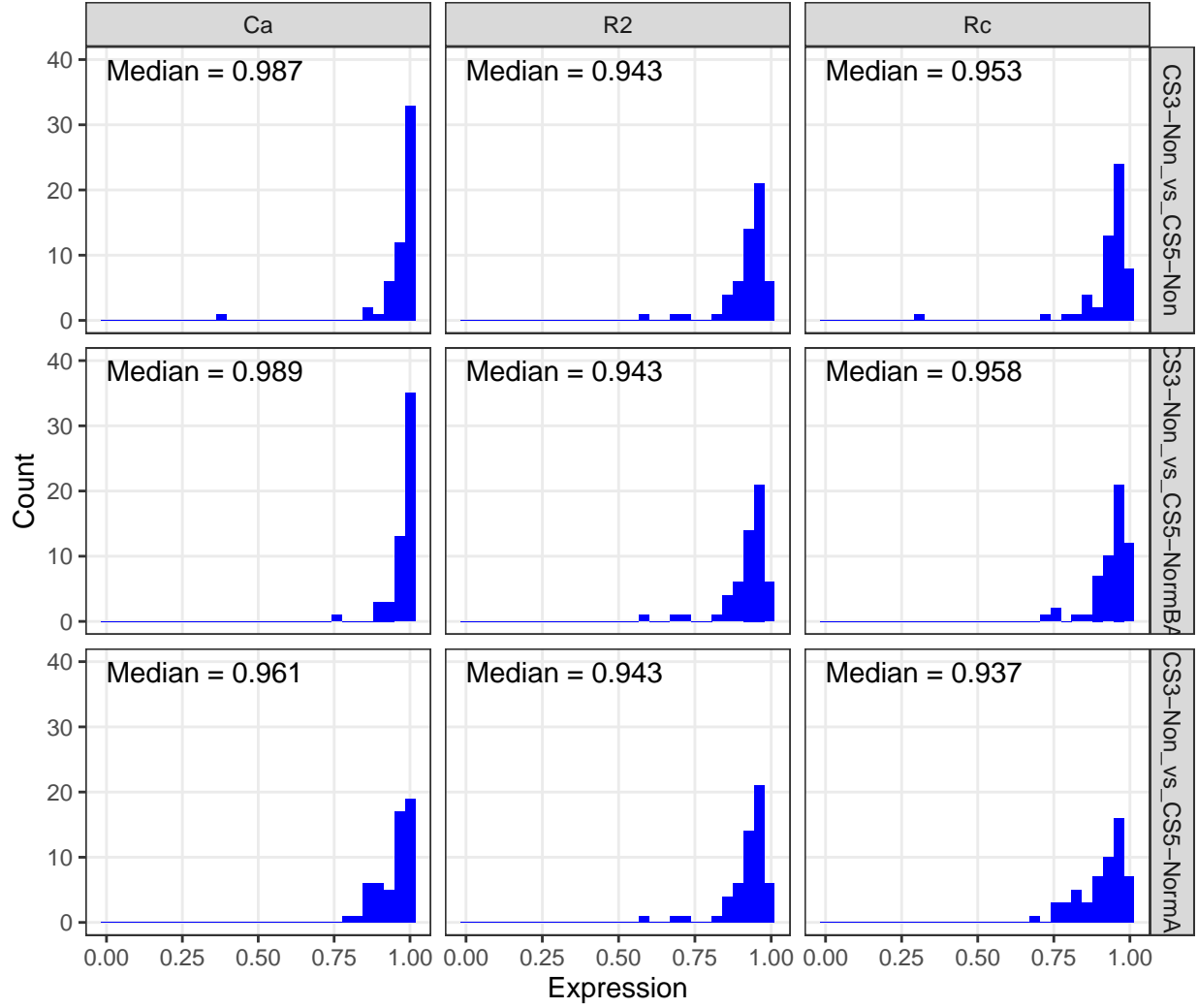


Figure 3.31: CS5 Set B/A Chaining Concordance Measure Distributions 2

CS4 Set A Chaining Concordance Measure Distributions

Samples=72, Genes=55

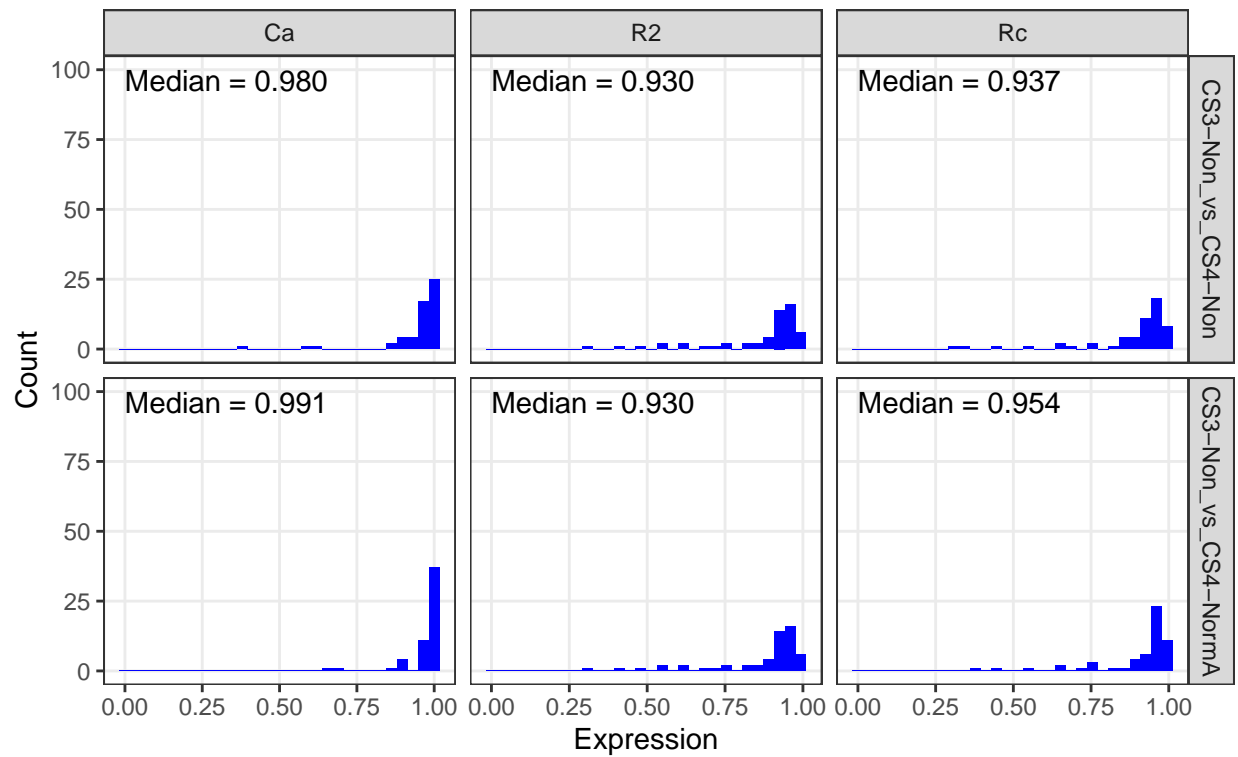


Figure 3.32: CS4 Set A Chaining Concordance Measure Distributions

CS4 and CS5 using Set B Concordance Measure Distributions

Samples=72, Genes=55

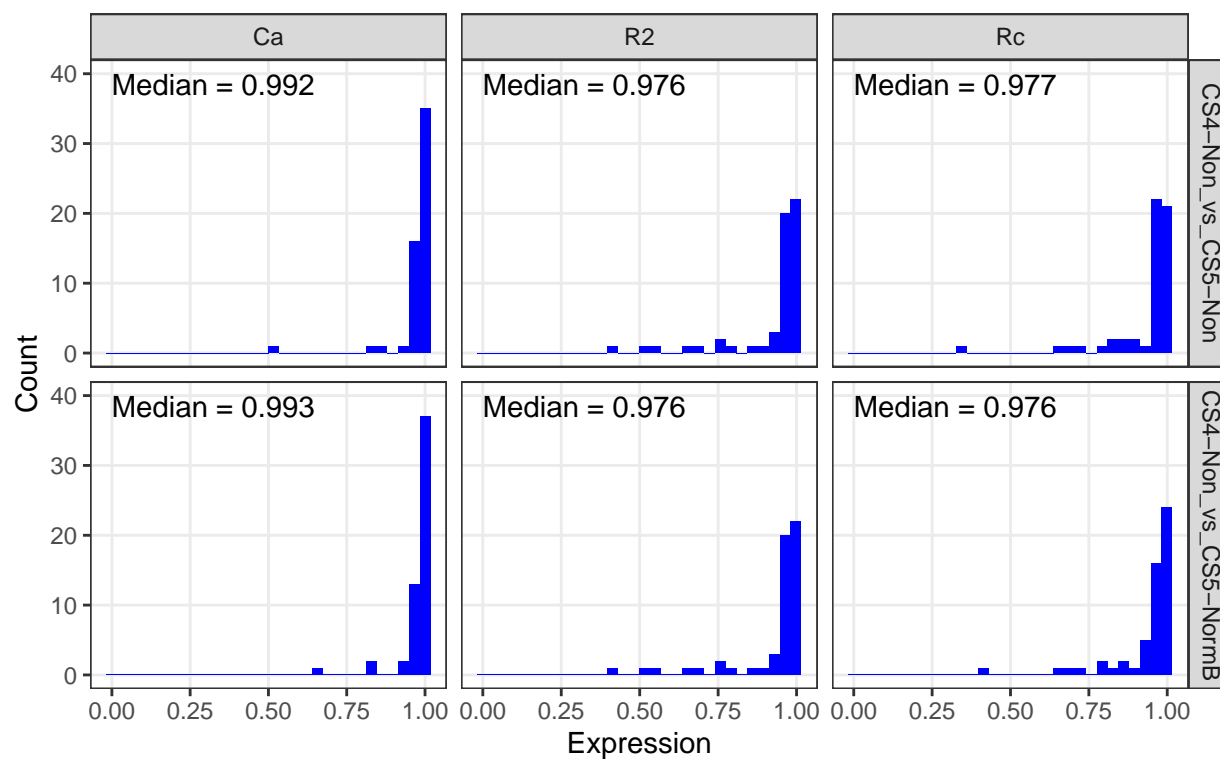


Figure 3.33: CS4 and CS5 using Set B Concordance Measure Distributions

3.3.4.3 CS3, CS4, CS5 using Set C/A

CS5 Set C/A Chaining Concordance Measure Distributions

Samples=72, Genes=55

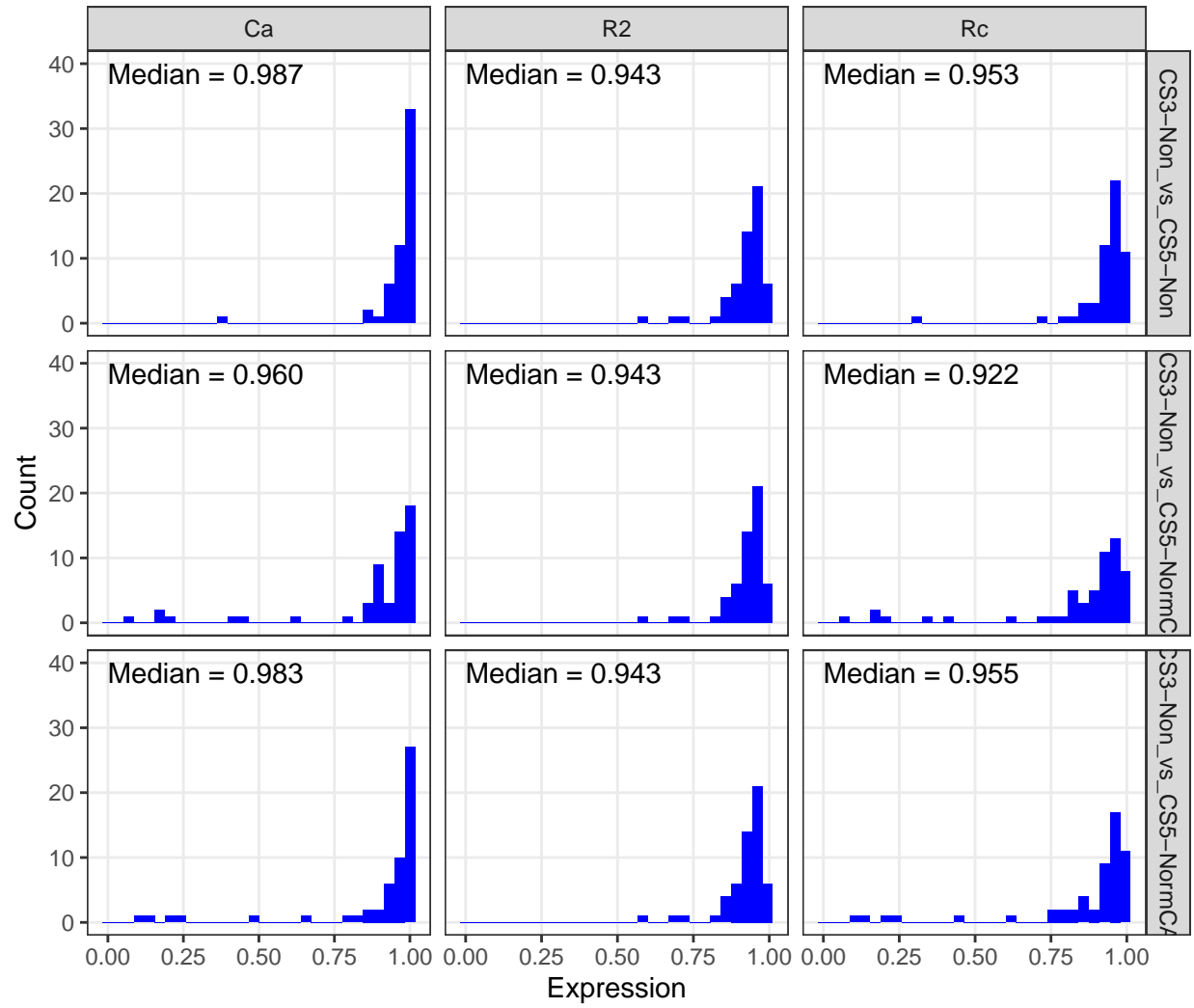


Figure 3.34: CS5 Set C/A Chaining Concordance Measure Distributions

Table 3.23: DSC for CS3 vs CS5 Set C/A Comparisons

| Comparisons | DSC | pval |
|-------------------|-------|-------|
| CS3 vs CS5-None | 0.126 | 0.101 |
| CS3 vs CS5-NormCA | 0.383 | 0.000 |
| CS3 vs CS5-NormA | 0.144 | 0.047 |

CS5 Set C/A Chaining Concordance Measure Distributions 2

Samples=72, Genes=55

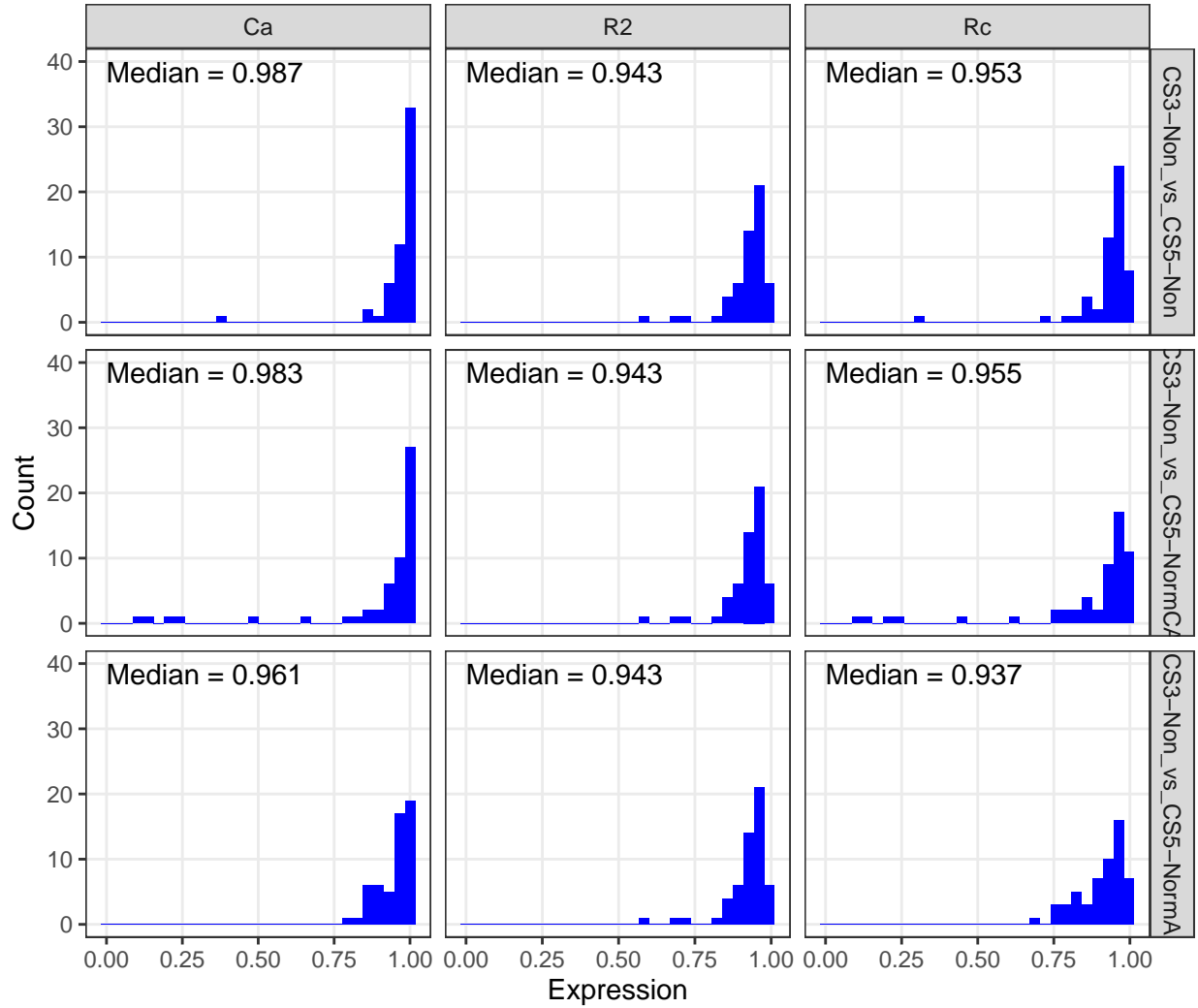


Figure 3.35: CS5 Set C/A Chaining Concordance Measure Distributions 2

Table 3.24: All Common Samples Histotype Distribution

| revHist | CS1 | CS2 | CS3 |
|---------|-----|-----|-----|
| CCOC | 3 | 4 | 3 |
| ENOC | 4 | 4 | 3 |
| HGSC | 59 | 62 | 75 |
| LGSC | 7 | 5 | 4 |
| MUC | 7 | 5 | 5 |

Table 3.25: Distinct Common Samples Histotype Distribution

| revHist | CS1 | CS2 | CS3 |
|---------|-----|-----|-----|
| CCOC | 3 | 3 | 3 |
| ENOC | 3 | 3 | 3 |
| HGSC | 57 | 57 | 57 |
| LGSC | 4 | 4 | 4 |
| MUC | 5 | 5 | 5 |

CS4 and CS5 using Set C Concordance Measure Distributions

Samples=72, Genes=55

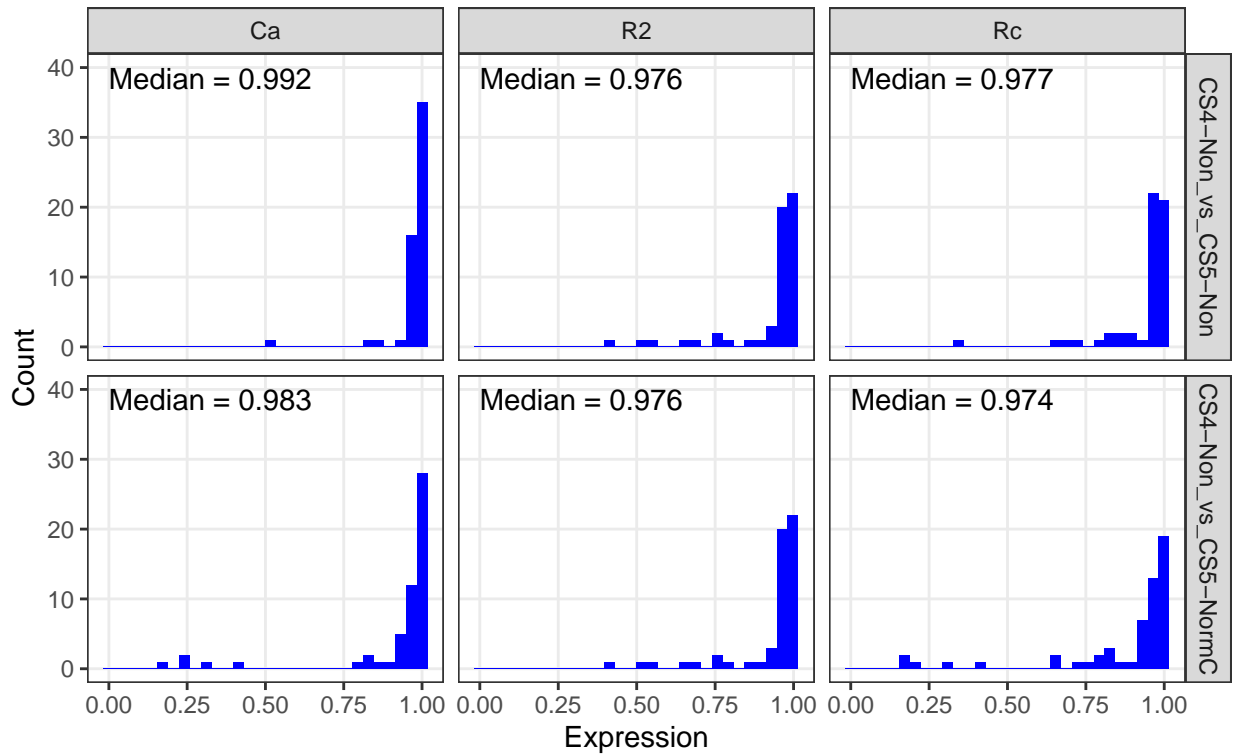


Figure 3.36: CS4 and CS5 using Set C Concordance Measure Distributions

Table 3.26: 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 3.27: 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 3.28: 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 3.29: CS3/CS4/CS5 Common Samples Histotype Distribution

| revHist | CS3 | CS4 | CS5 |
|---------|-----|-----|-----|
| HGSC | 46 | 46 | 46 |
| NA | 26 | 26 | 26 |

Table 3.30: 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 3.31: Full Training Set Histotype Distribution

| revHist | n | freq |
|---------|------|------|
| HGSC | 1227 | 79% |
| CCOC | 106 | 7% |
| ENOC | 91 | 6% |
| MUC | 84 | 5% |
| LGSC | 39 | 3% |

Table 3.32: 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%) |

3.4 Common Sample Distributions

3.5 Histotype Distribution in Classifier Datasets

Table 3.33: CS1 All Training Set Histotype Distribution

| revHist | n | freq |
|---------|-----|------|
| HGSC | 125 | 47% |
| ENOC | 58 | 22% |
| CCOC | 47 | 18% |
| LGSC | 19 | 7% |
| MUC | 19 | 7% |

Table 3.34: CS2 All Training Set Histotype Distribution

| revHist | n | freq |
|---------|-----|------|
| HGSC | 654 | 79% |
| MUC | 61 | 7% |
| CCOC | 60 | 7% |
| ENOC | 32 | 4% |
| LGSC | 20 | 2% |

Table 3.35: Confirmation Set Histotype Distribution

| revHist | n | freq |
|---------|-----|------|
| HGSC | 423 | 66% |
| ENOC | 106 | 16% |
| CCOC | 75 | 12% |
| MUC | 27 | 4% |
| LGSC | 13 | 2% |

Table 3.36: Validation Set Histotype Distribution

| revHist | n | freq |
|---------|-----|------|
| HGSC | 781 | 74% |
| ENOC | 140 | 13% |
| CCOC | 86 | 8% |
| MUC | 34 | 3% |
| LGSC | 20 | 2% |

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 overall accuracy 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

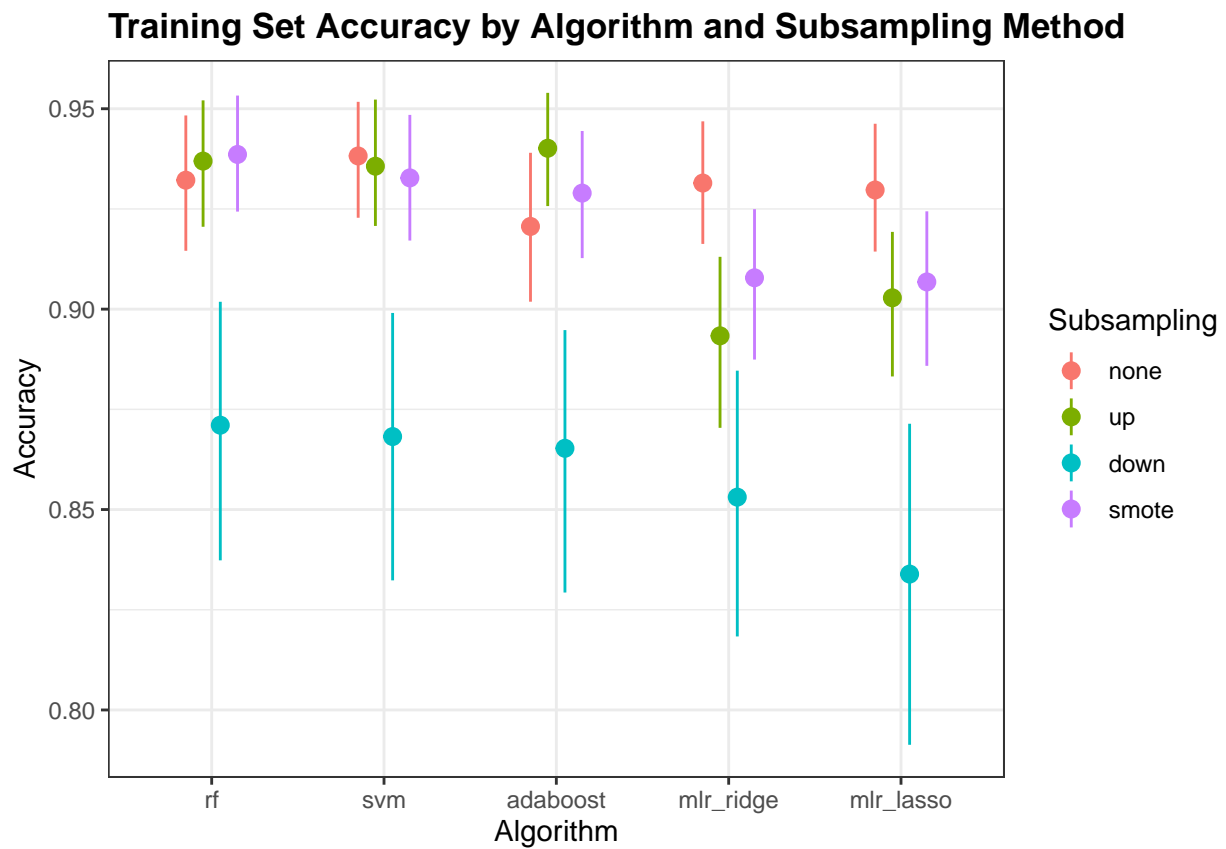


Figure 4.1: Training Set Accuracy

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 |

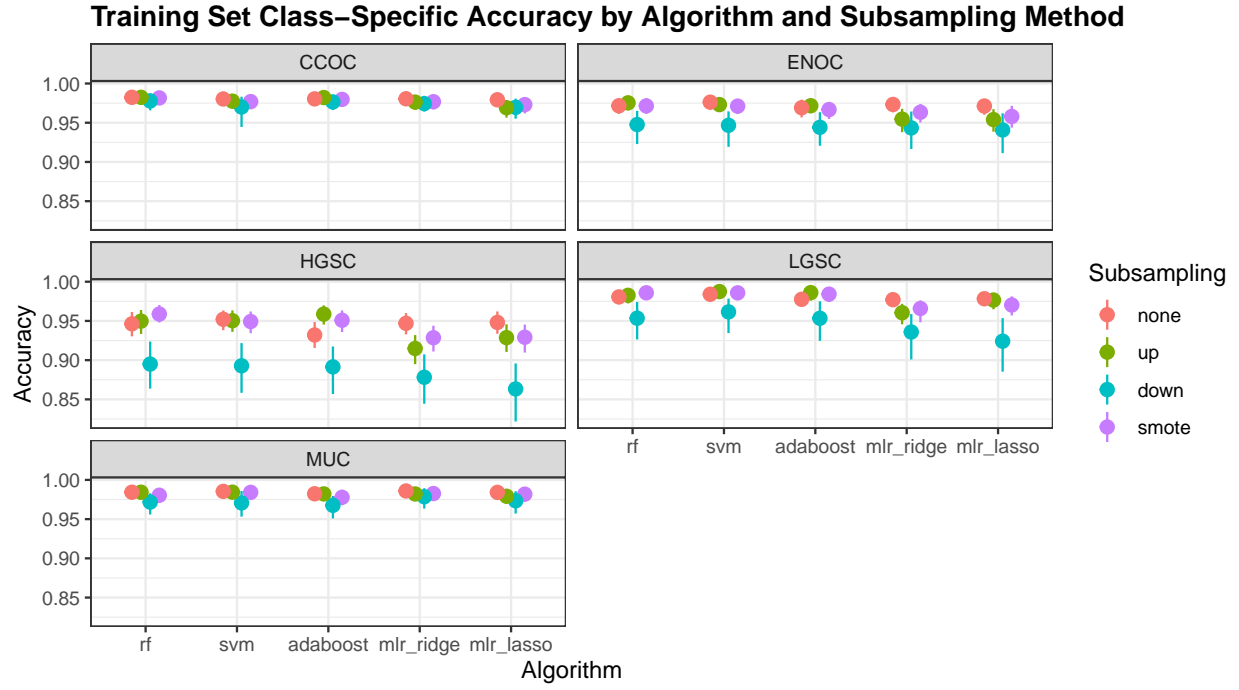


Figure 4.2: Training Set Class-Specific Accuracy

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 |

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 |

4.1.2 F1-Score

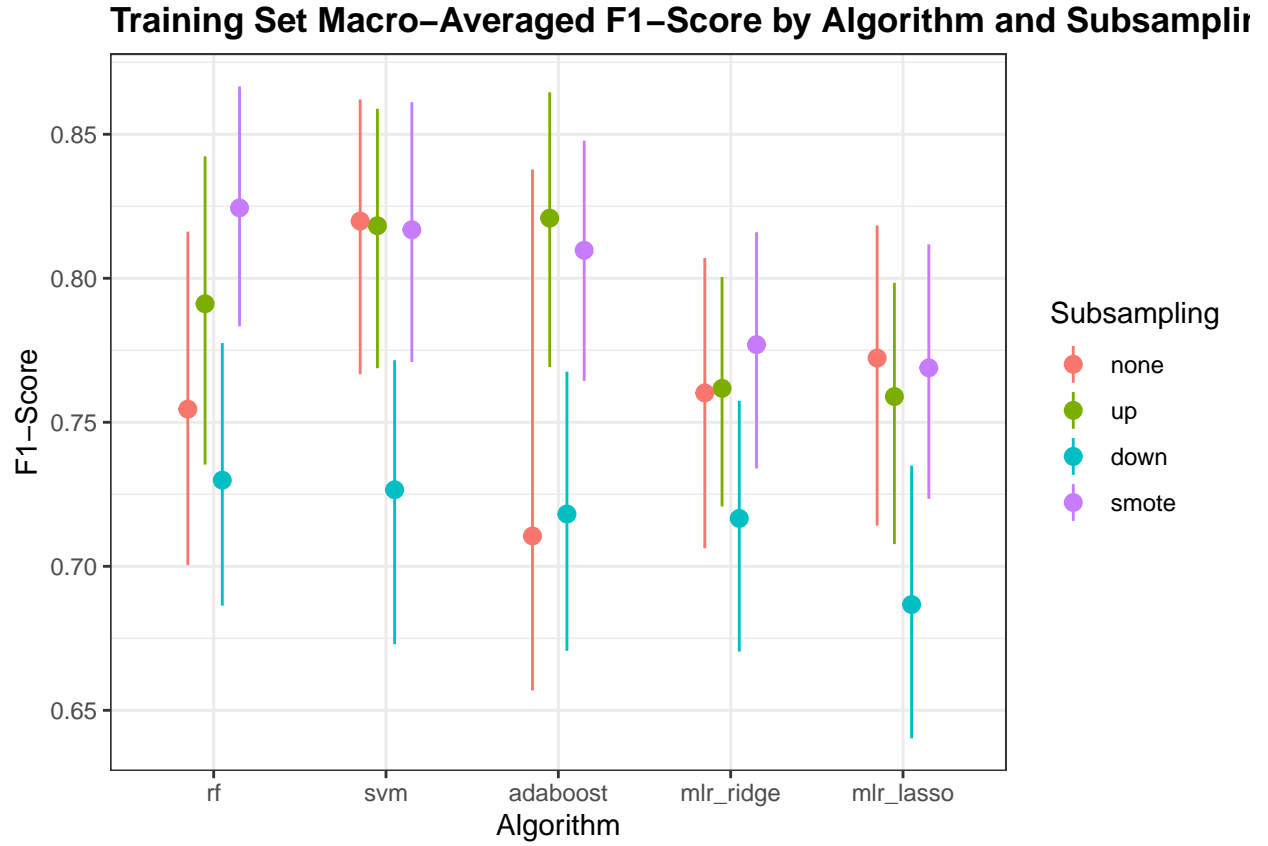


Figure 4.3: Training Set F1-Score

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 | 0.866 | 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 |

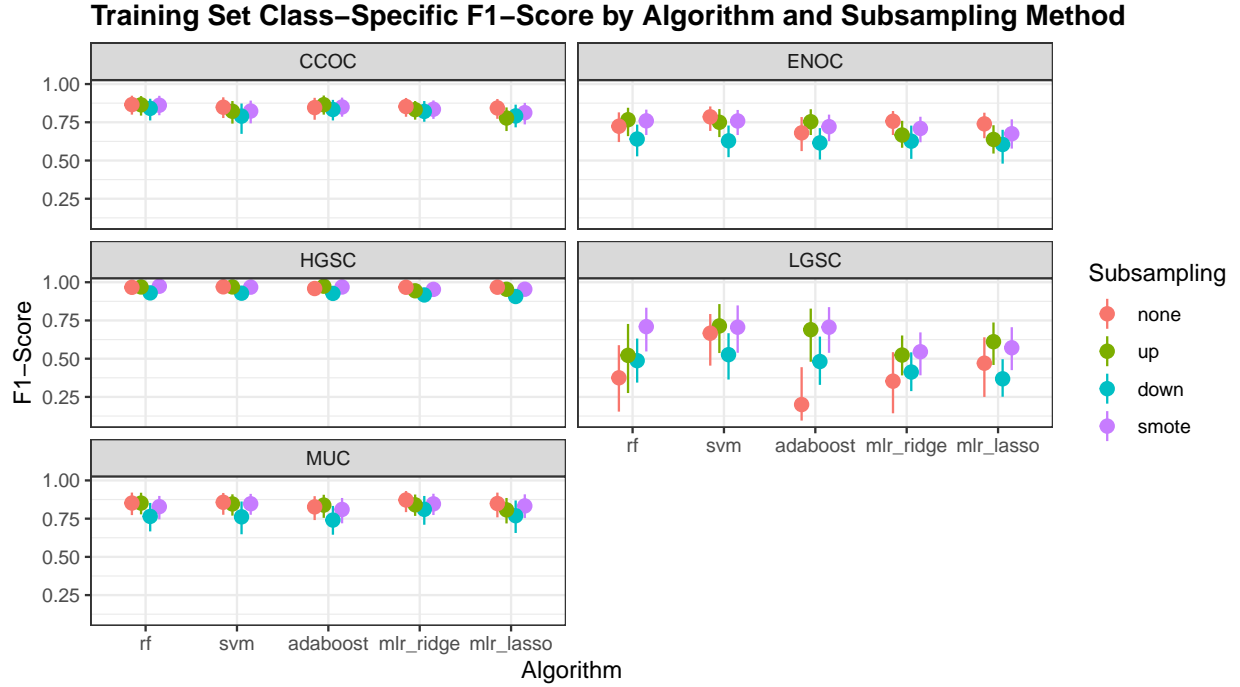


Figure 4.4: Training Set Class-Specific F1-Score

Table 4.5: Training Set Kappa by Algorithm and Subsampling Method

| sampling | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|----------|-------------|-------|----------|-----------|-----------|
| none | 0.793 | 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 |

4.1.3 Kappa

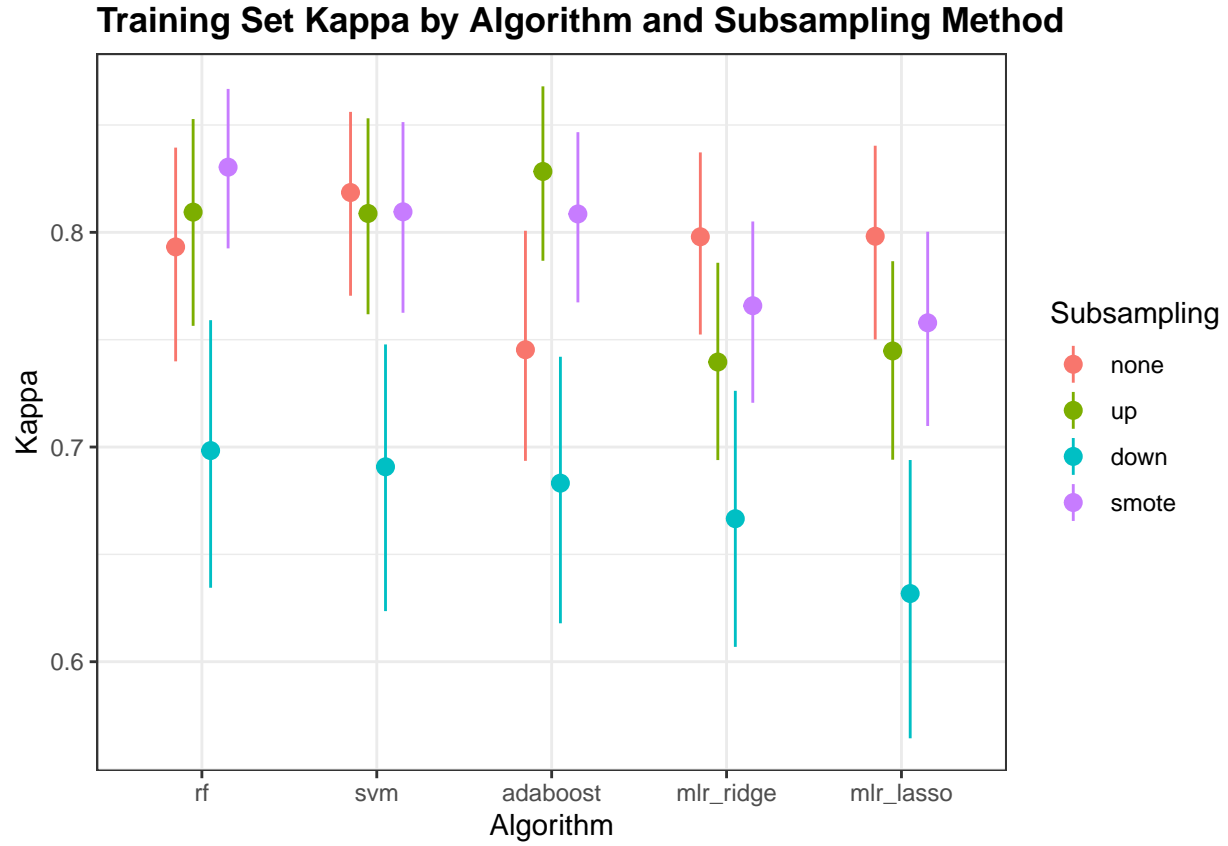


Figure 4.5: Training Set Kappa

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

| sampling | histotype | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|----------|-----------|--------------|-------|----------|-----------|-----------|
| none | CCOC | 0.857 | 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 |

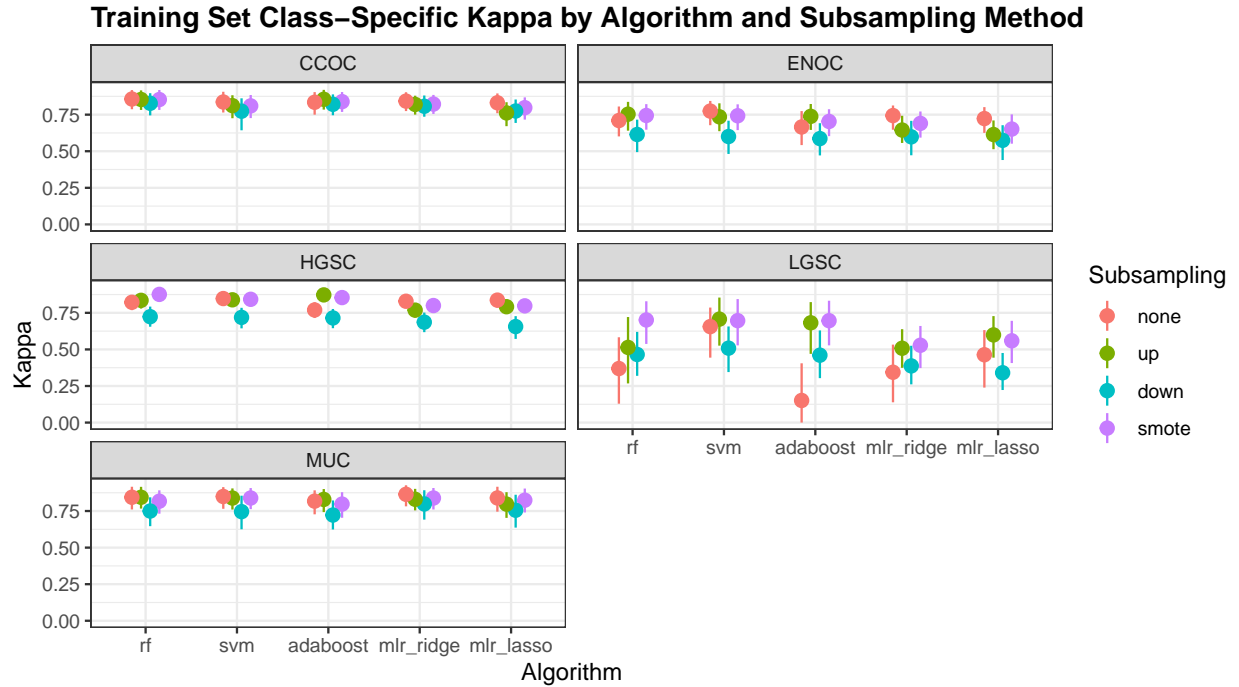


Figure 4.6: Training Set Class-Specific Kappa

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

| sampling | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|----------|-------|-------|----------|--------------|-----------|
| none | 0.632 | 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 |

4.1.4 G-mean

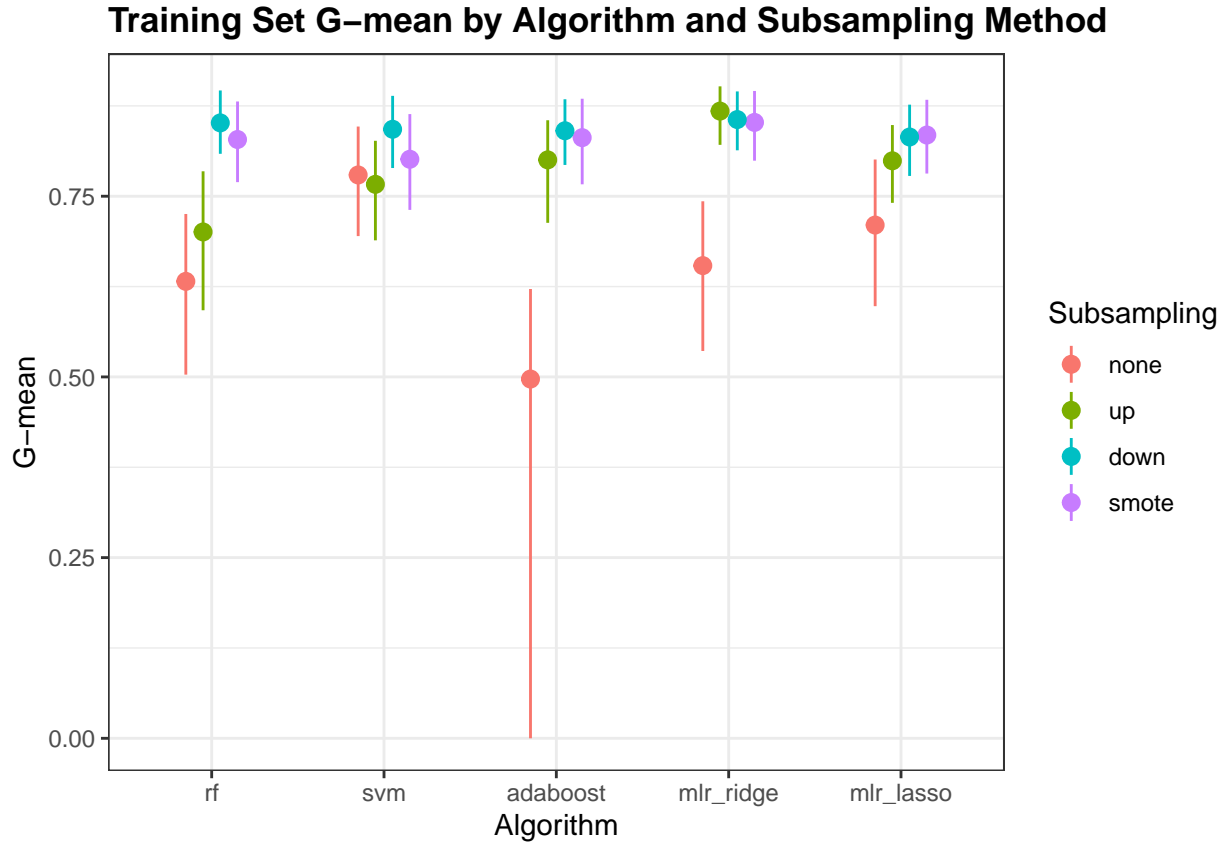


Figure 4.7: Training Set G-mean

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 | 0.896 | 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 |

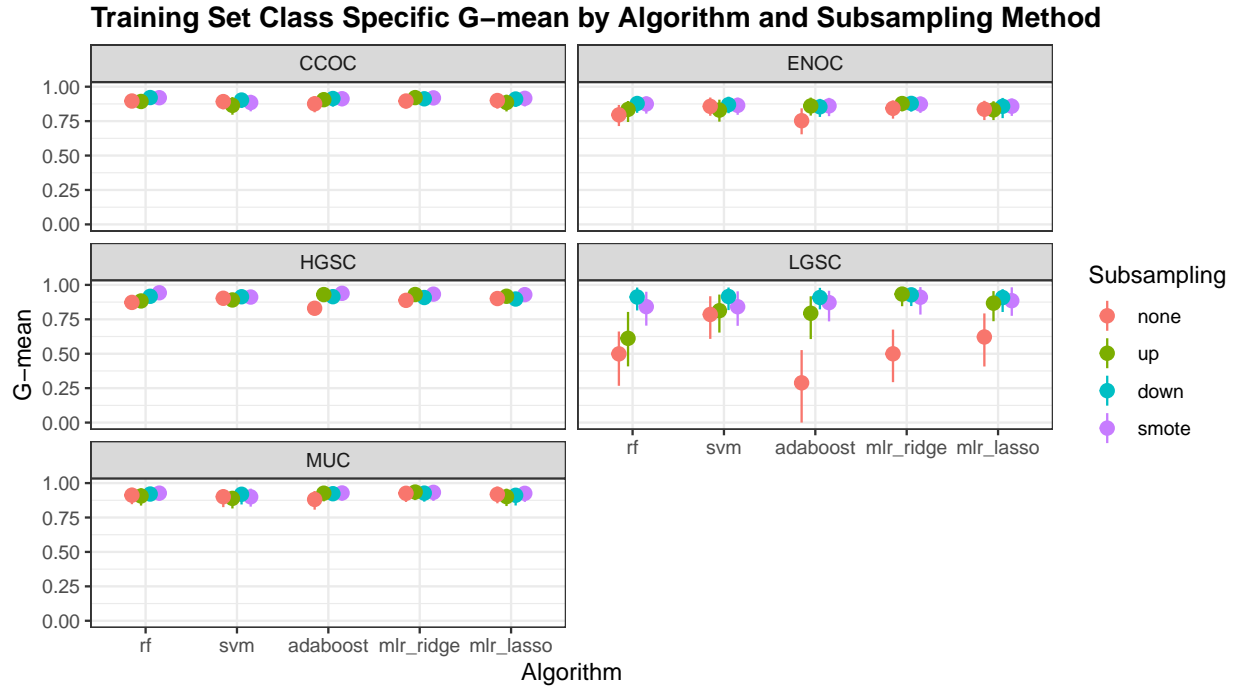


Figure 4.8: Training Set Class-Specific G-mean

4.2 Two-Step Training Set

4.2.1 Accuracy

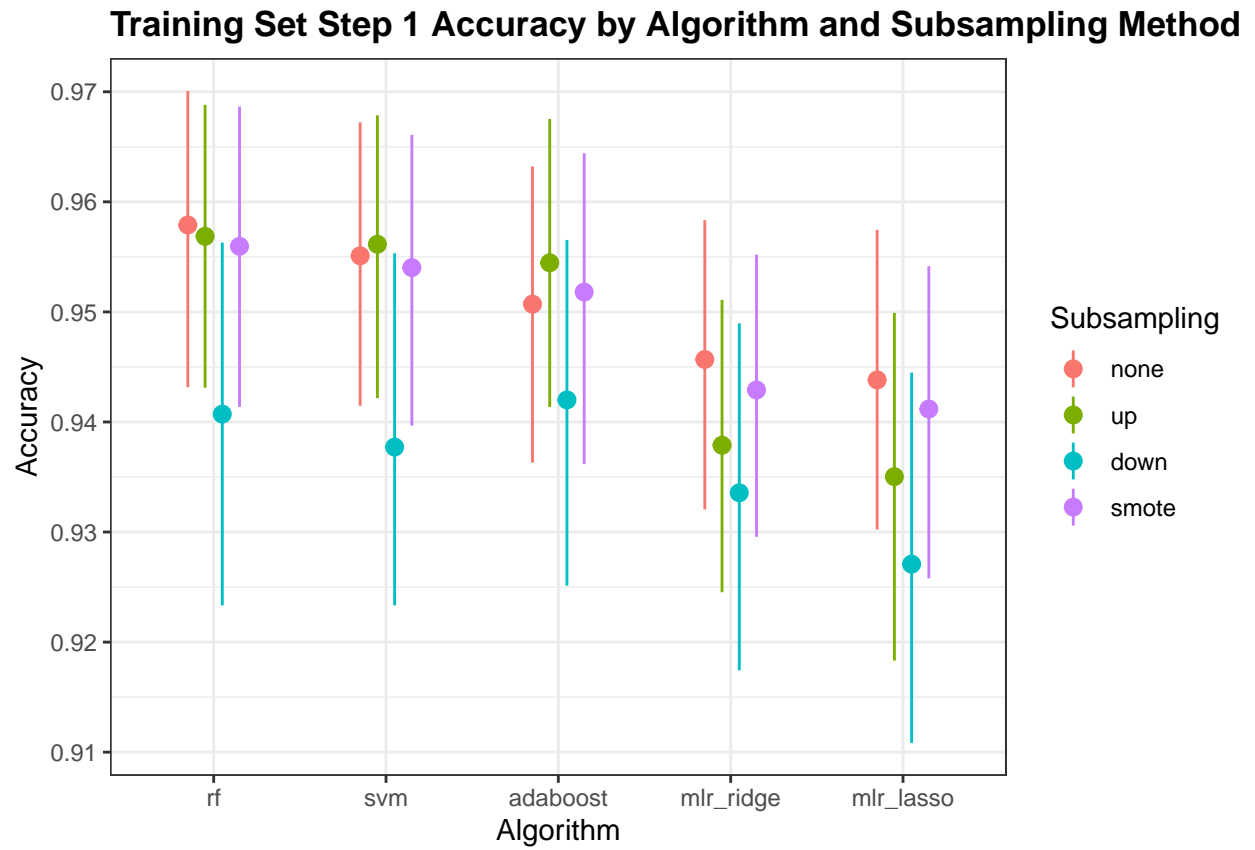


Figure 4.9: Training Set Step 1 Accuracy

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

| sampling | rf | svm | adaboost | mlr_ridge | mlr_lasso |
|----------|--------------|-------|----------|-----------|-----------|
| none | 0.958 | 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 |

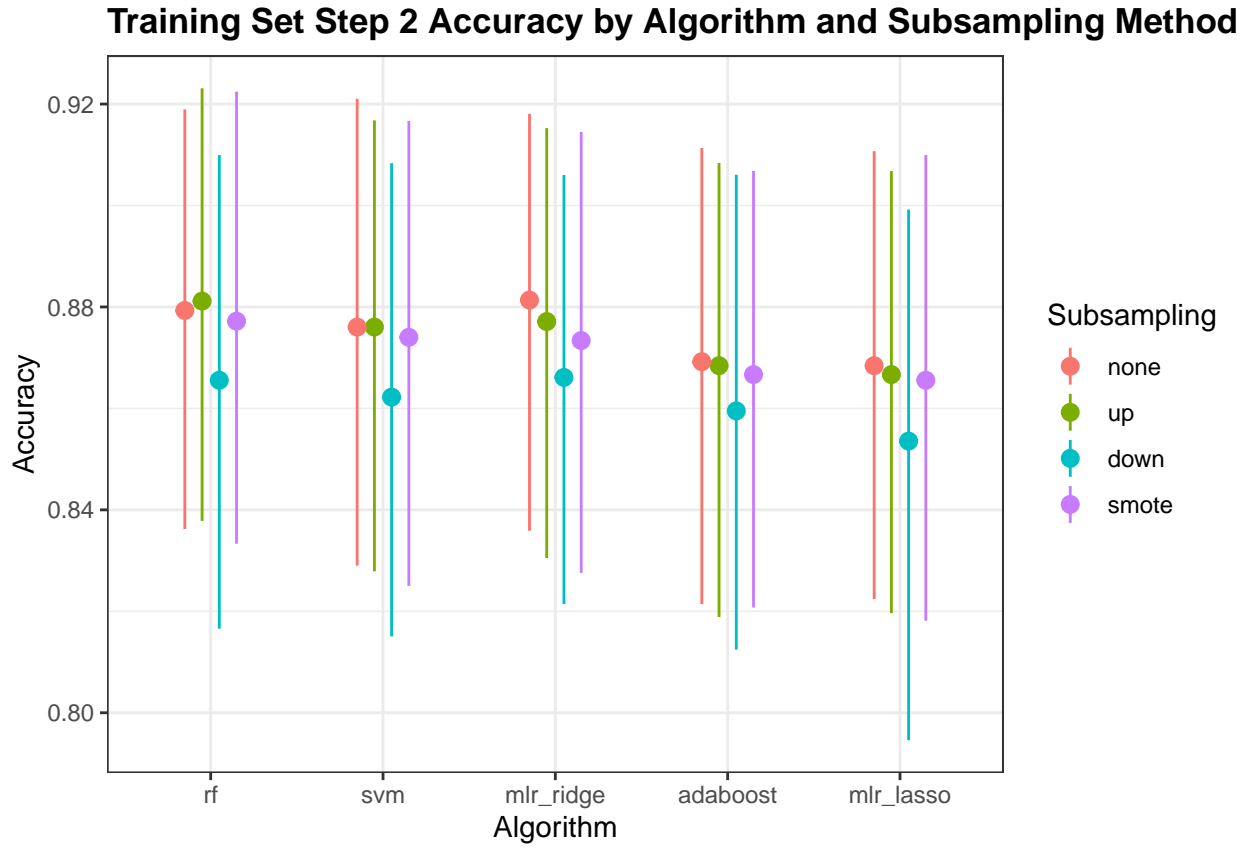


Figure 4.10: Training Set Step 2 Accuracy

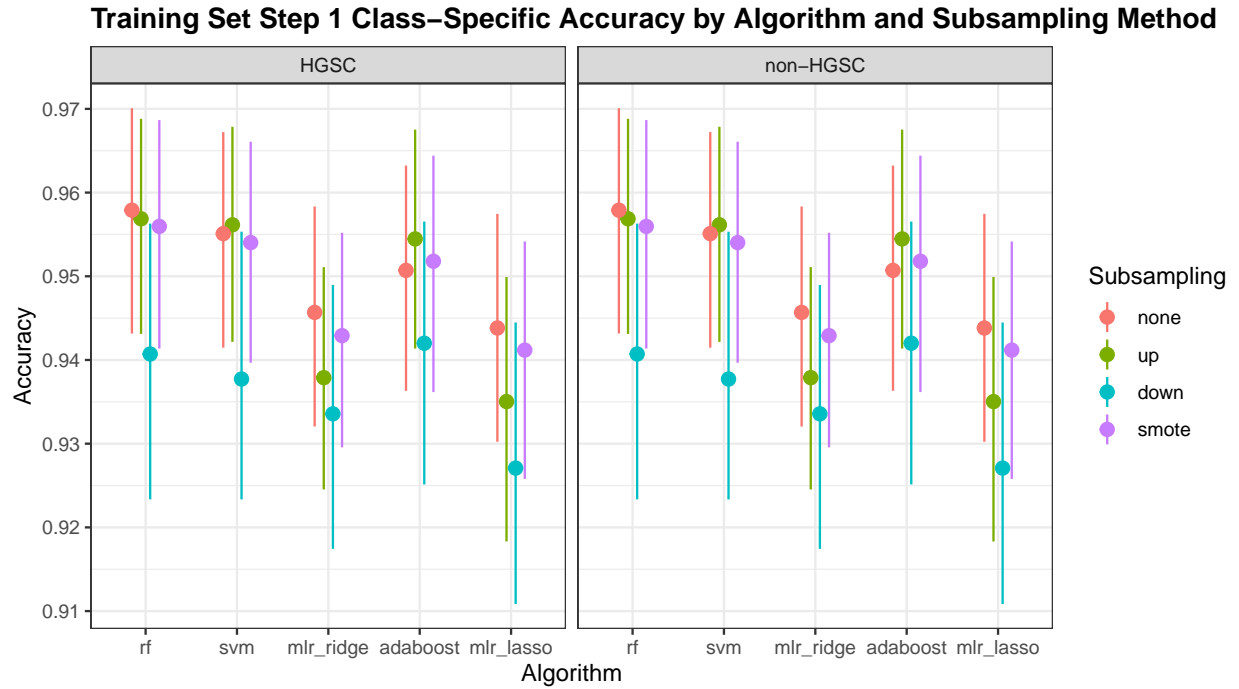


Figure 4.11: Training Set Step 1 Class-Specific Accuracy

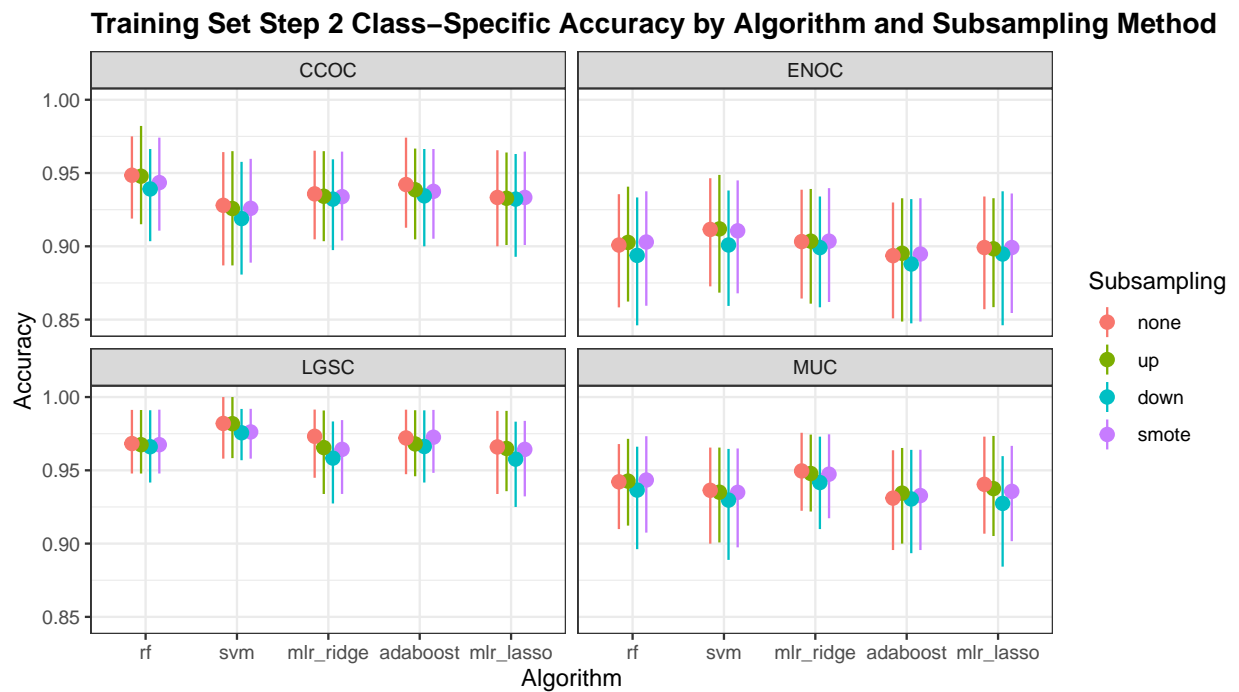


Figure 4.12: Training Set Step 2 Class-Specific Accuracy

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

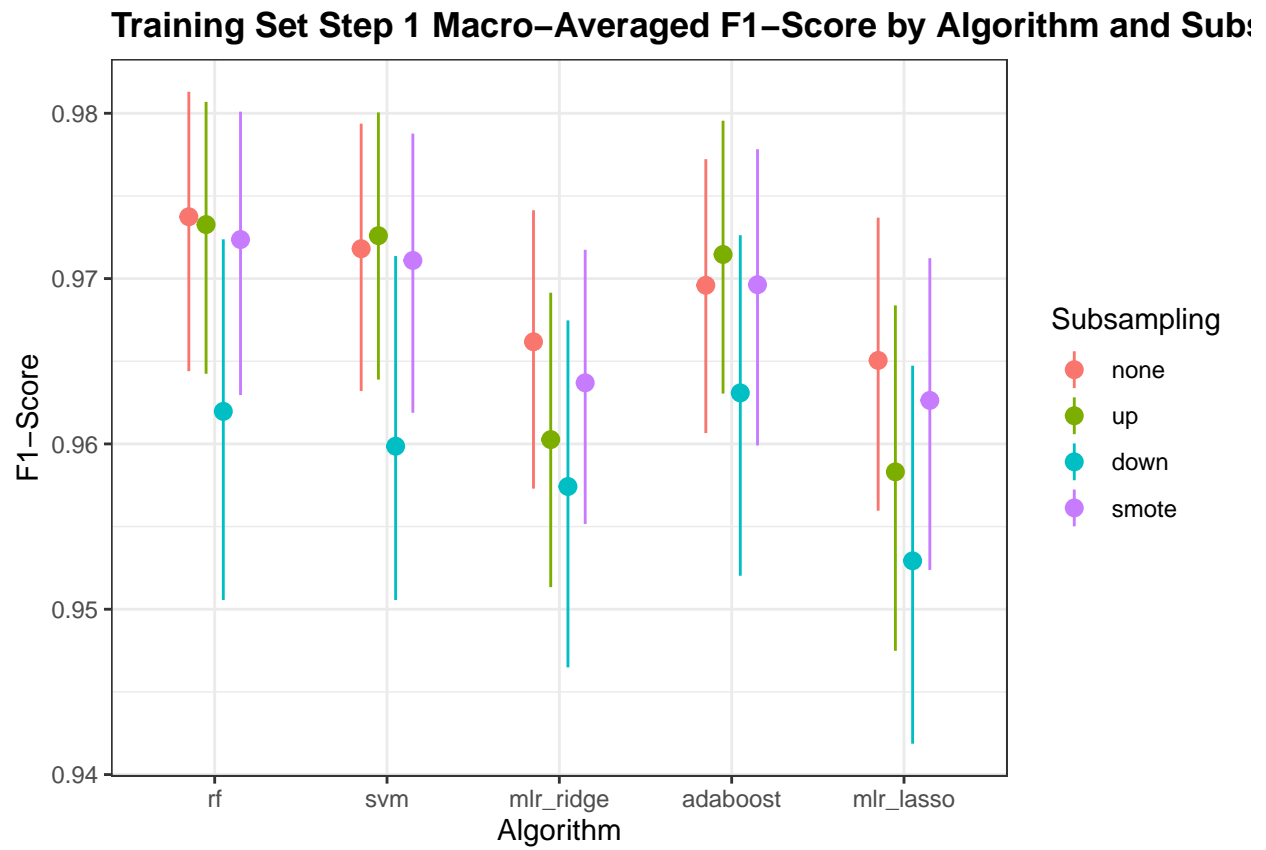


Figure 4.13: Training Set Step 1 F1-Score

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 |

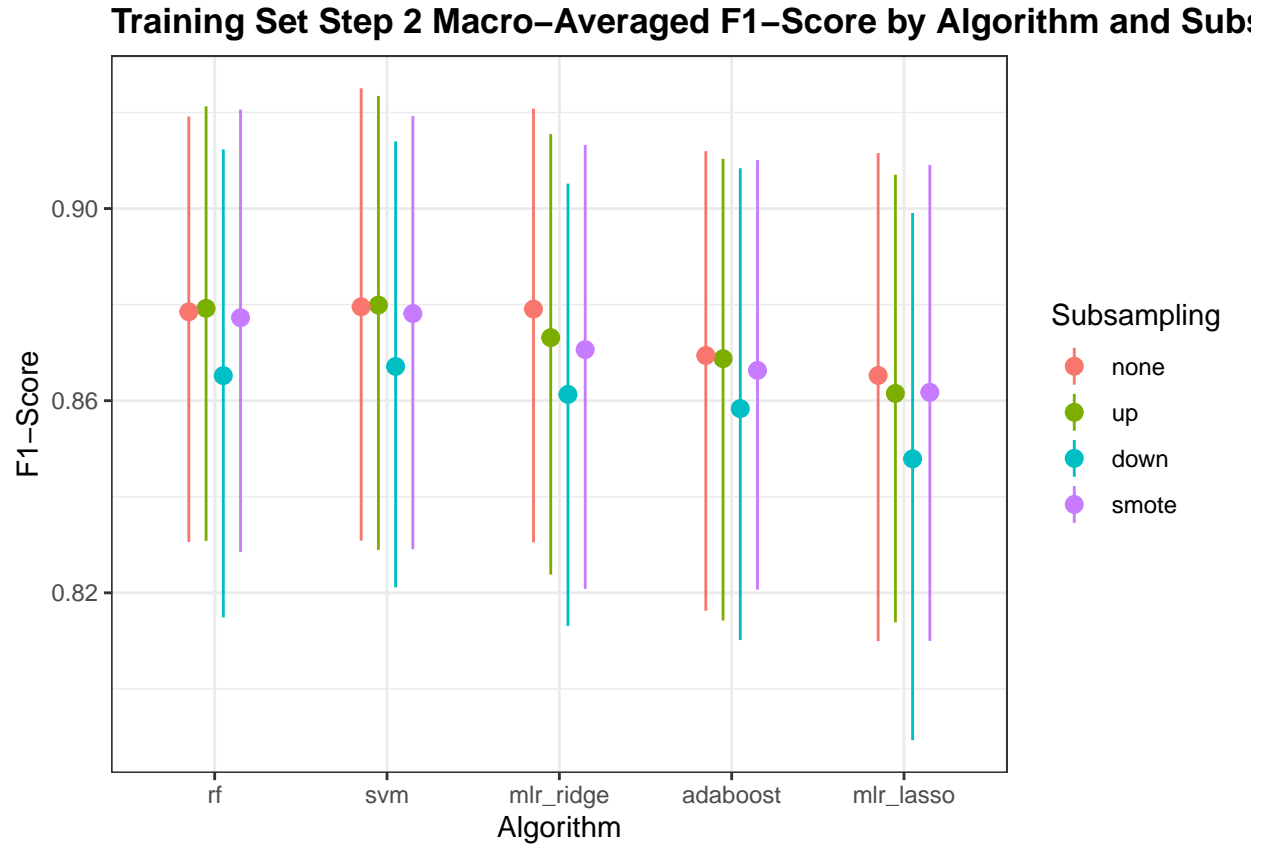


Figure 4.14: Training Set Step 2 F1-Score

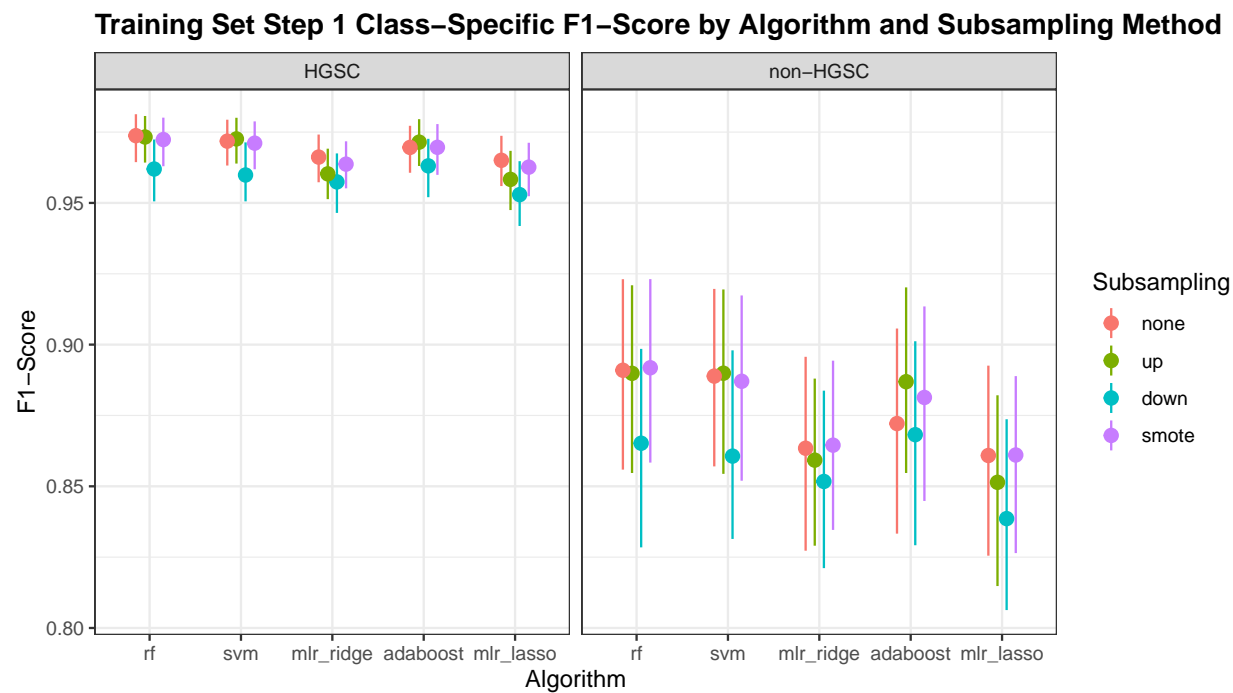


Figure 4.15: Training Set Step 1 Class-Specific F1-Score

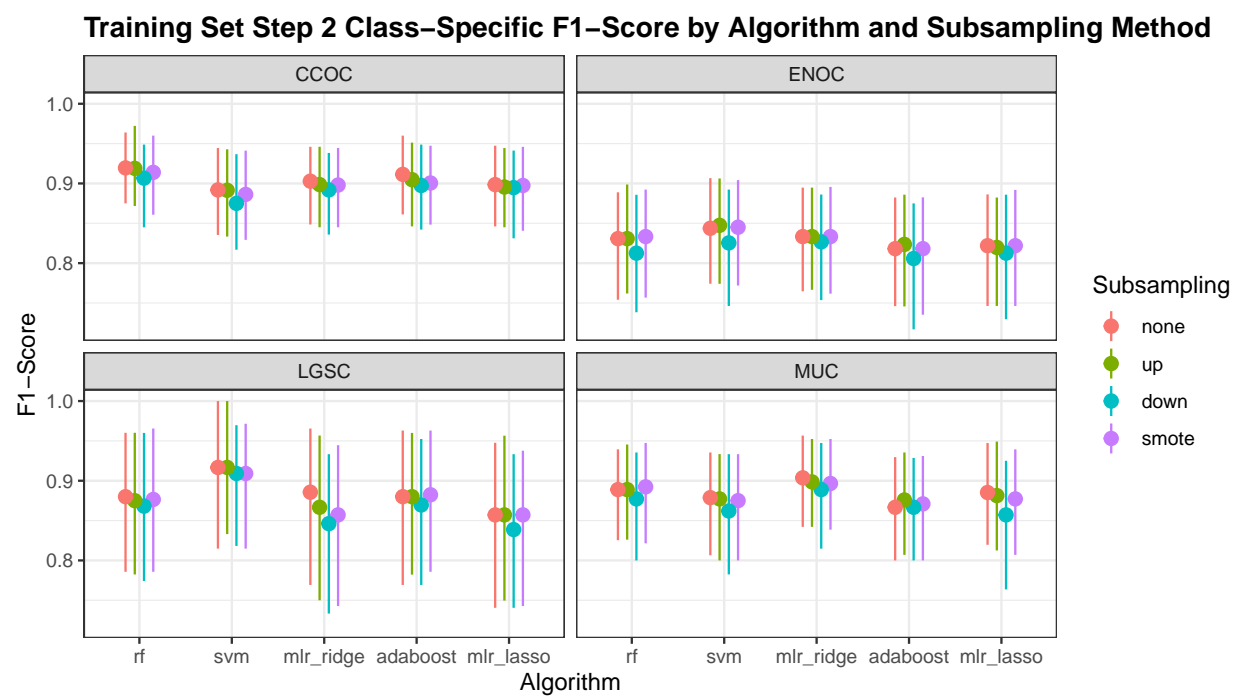


Figure 4.16: Training Set Step 2 Class-Specific F1-Score

Table 4.15: Training Set Step 1 Class-Specific F1-Score by Algorithm and Subsampling Method

| sampling | histotype | rf | svm | mlr_ridge | adaboost | mlr_lasso |
|----------|-----------|--------------|-------|-----------|----------|-----------|
| none | HGSC | 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 |

Table 4.16: Training Set Step 2 Class-Specific F1-Score by Algorithm and Subsampling Method

| sampling | histotype | rf | svm | mlr_ridge | adaboost | mlr_lasso |
|----------|-----------|-------------|-------|-----------|----------|-----------|
| none | CCOC | 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 |

4.2.3 Kappa

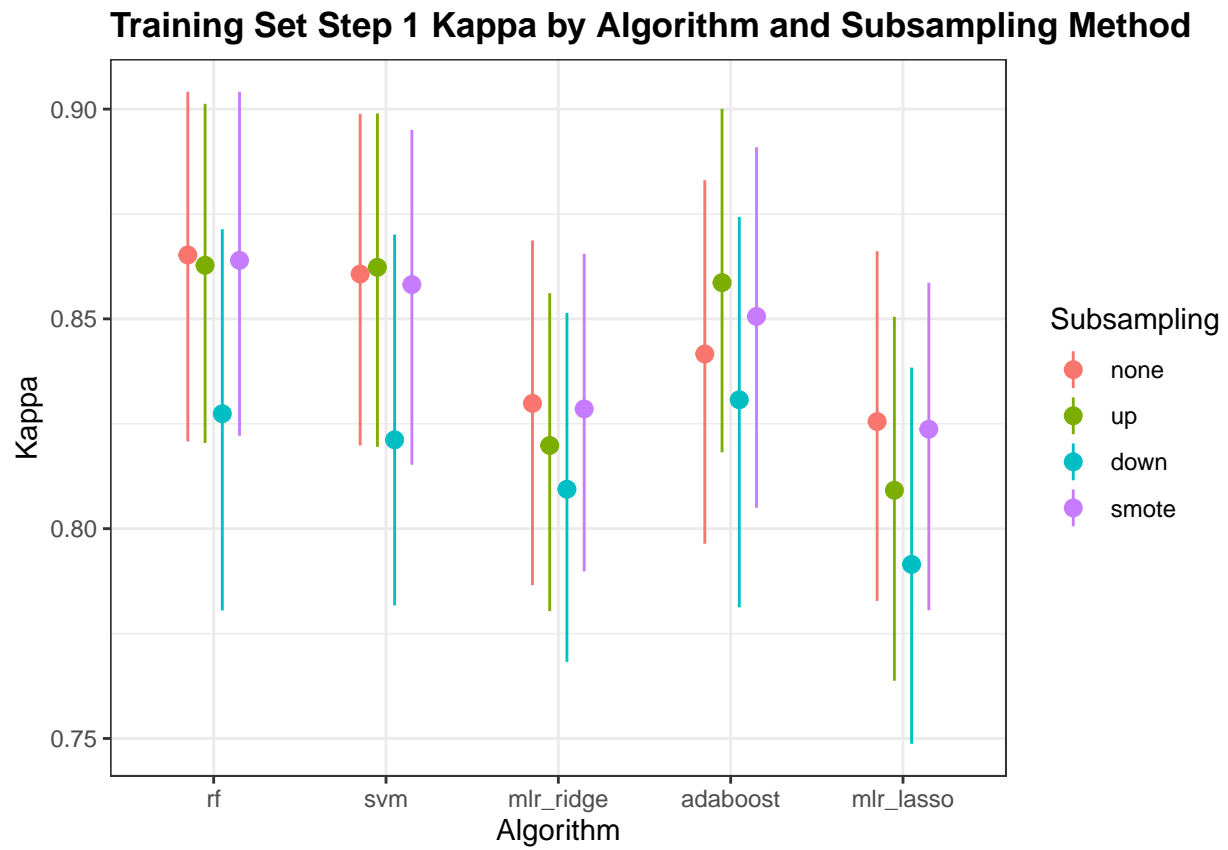


Figure 4.17: Training Set Step 1 Kappa

Table 4.17: 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 4.18: 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 |

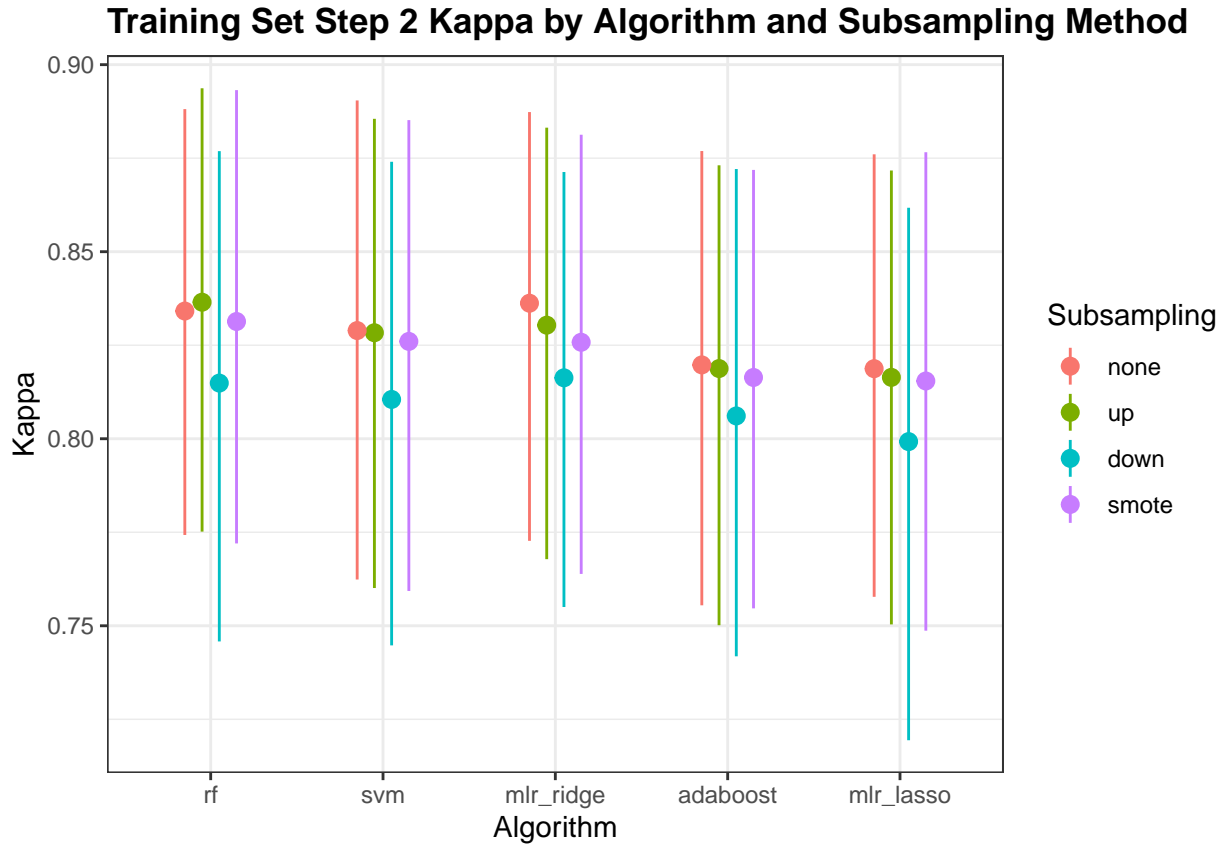


Figure 4.18: Training Set Step 2 Kappa

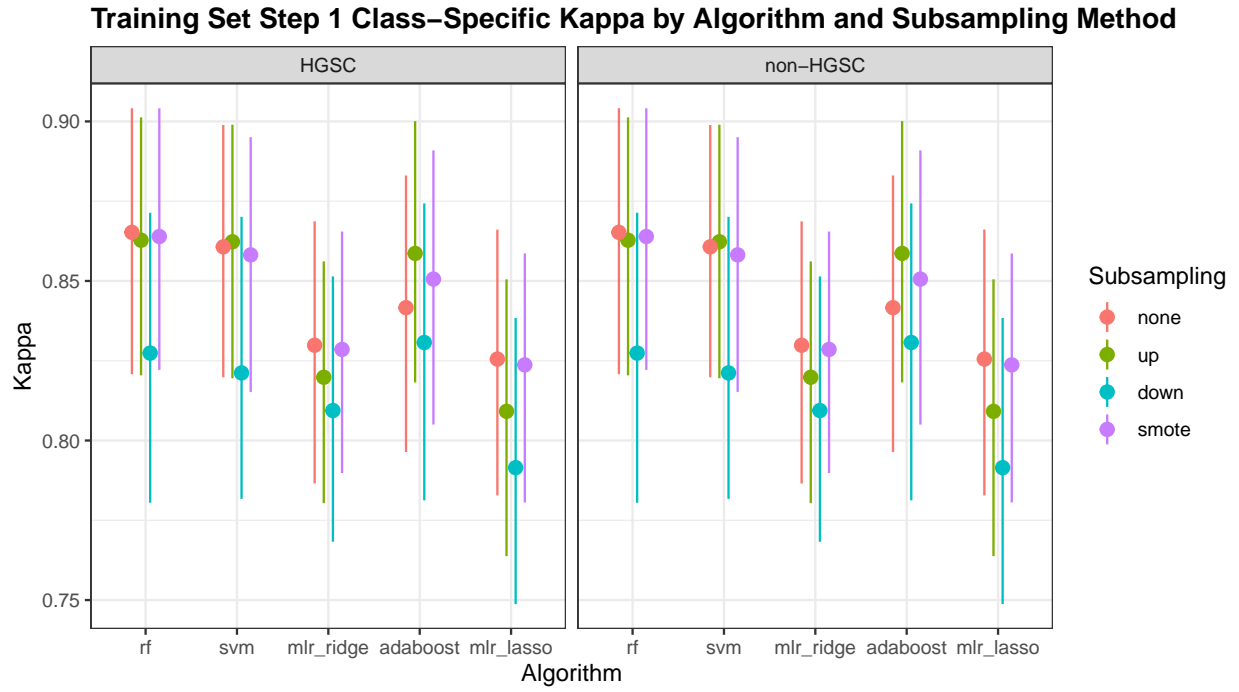


Figure 4.19: Training Set Step 1 Class-Specific Kappa

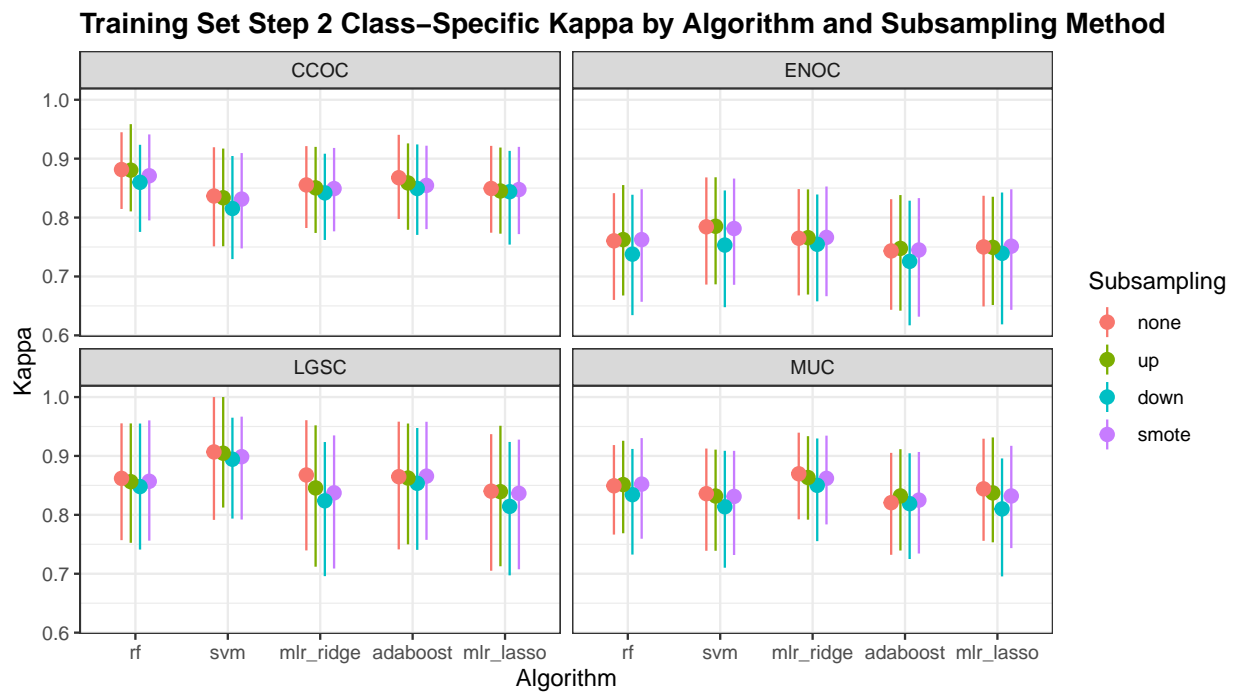


Figure 4.20: Training Set Step 2 Class-Specific Kappa

Table 4.19: 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 4.20: 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

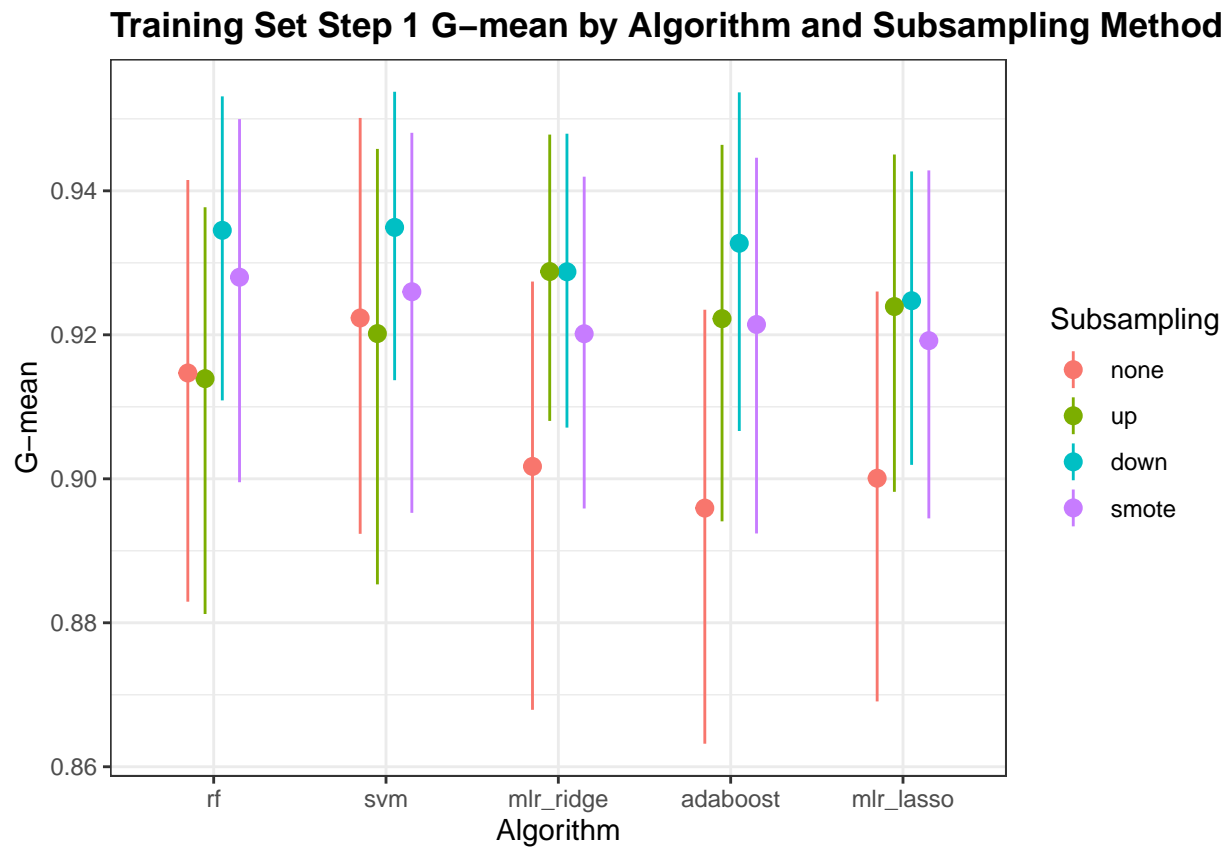


Figure 4.21: Training Set Step 1 G-mean

Table 4.21: 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 4.22: 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 |

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

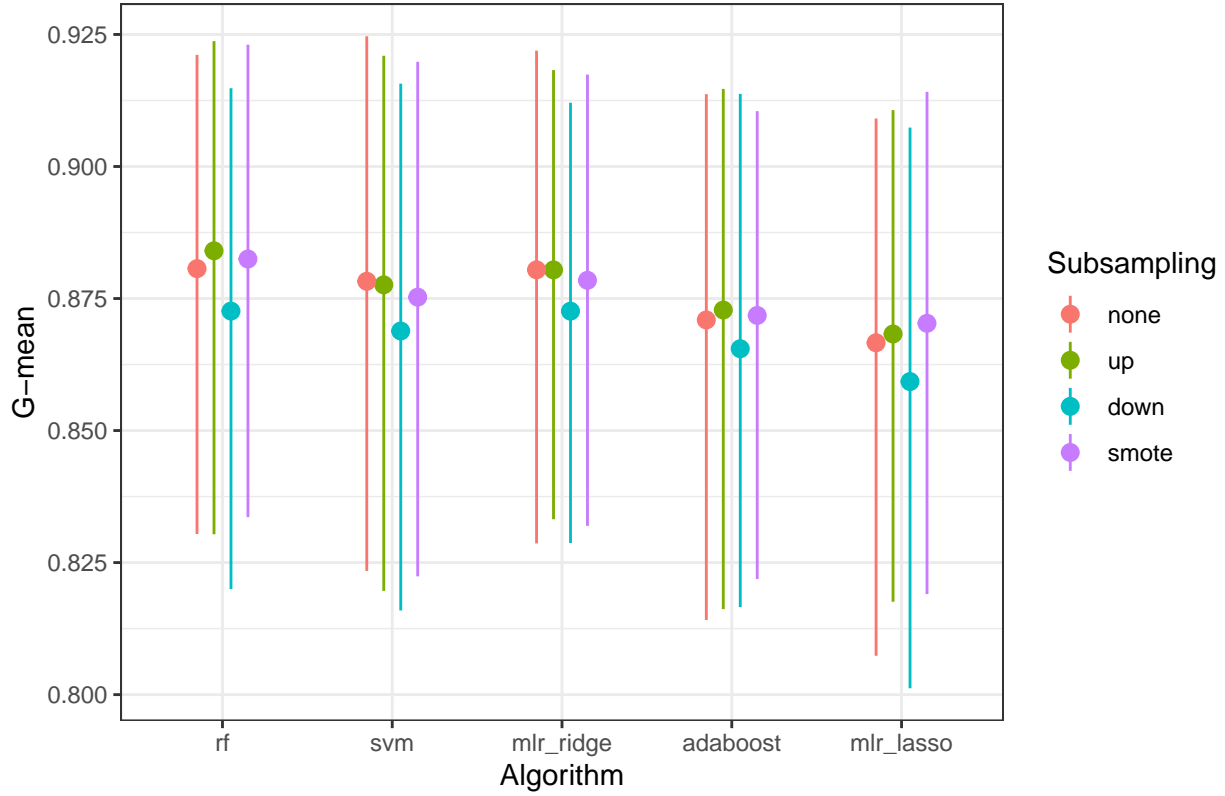


Figure 4.22: Training Set Step 2 G-mean

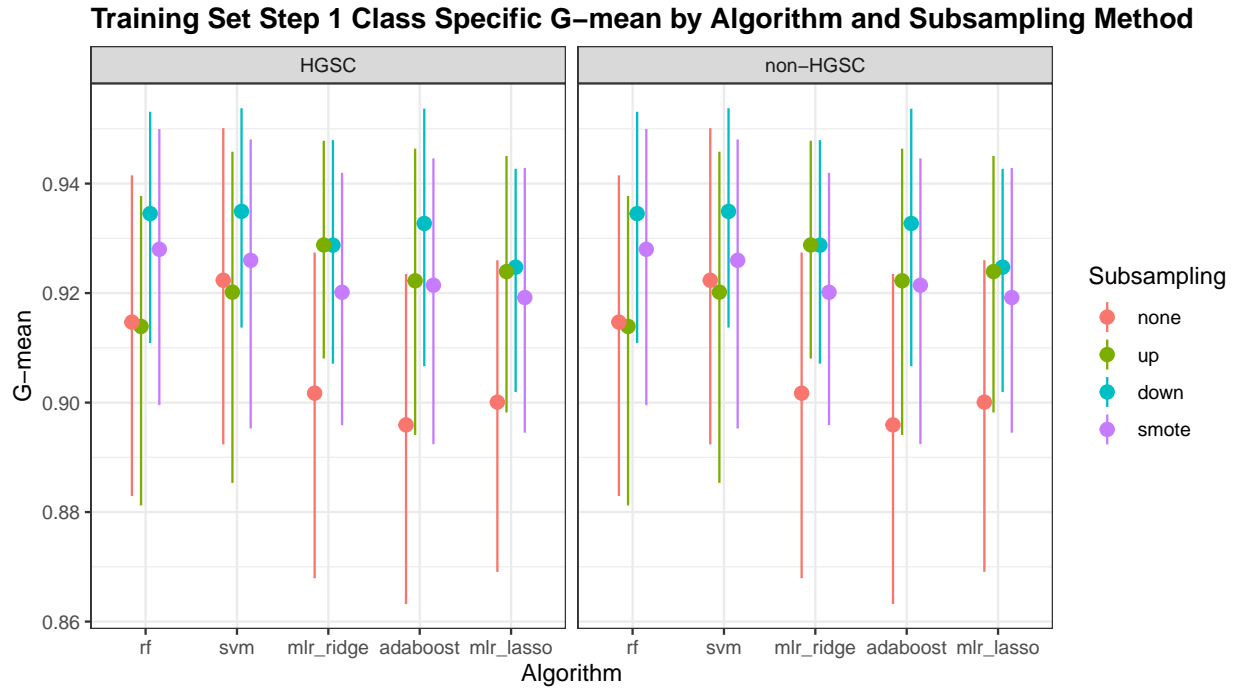


Figure 4.23: Training Set Step 1 Class-Specific G-mean

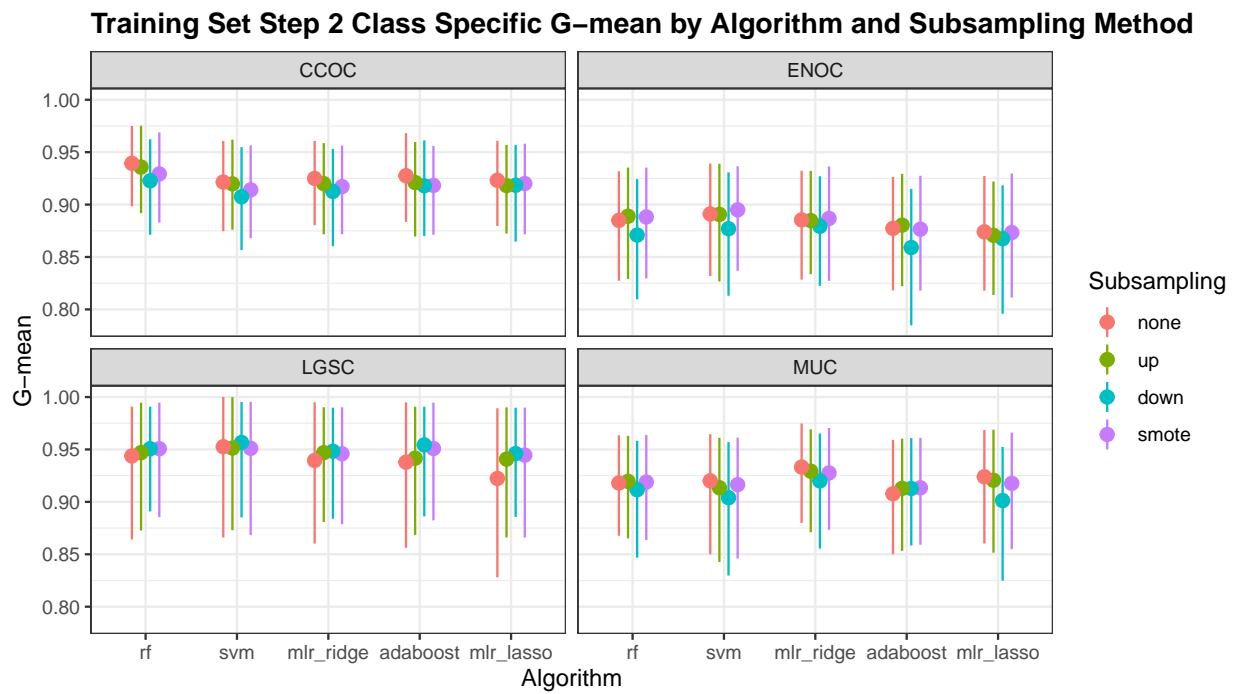


Figure 4.24: Training Set Step 2 Class-Specific G-mean

Table 4.23: 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 4.24: 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 4.25: 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 CS1 Set

4.3.1 Accuracy

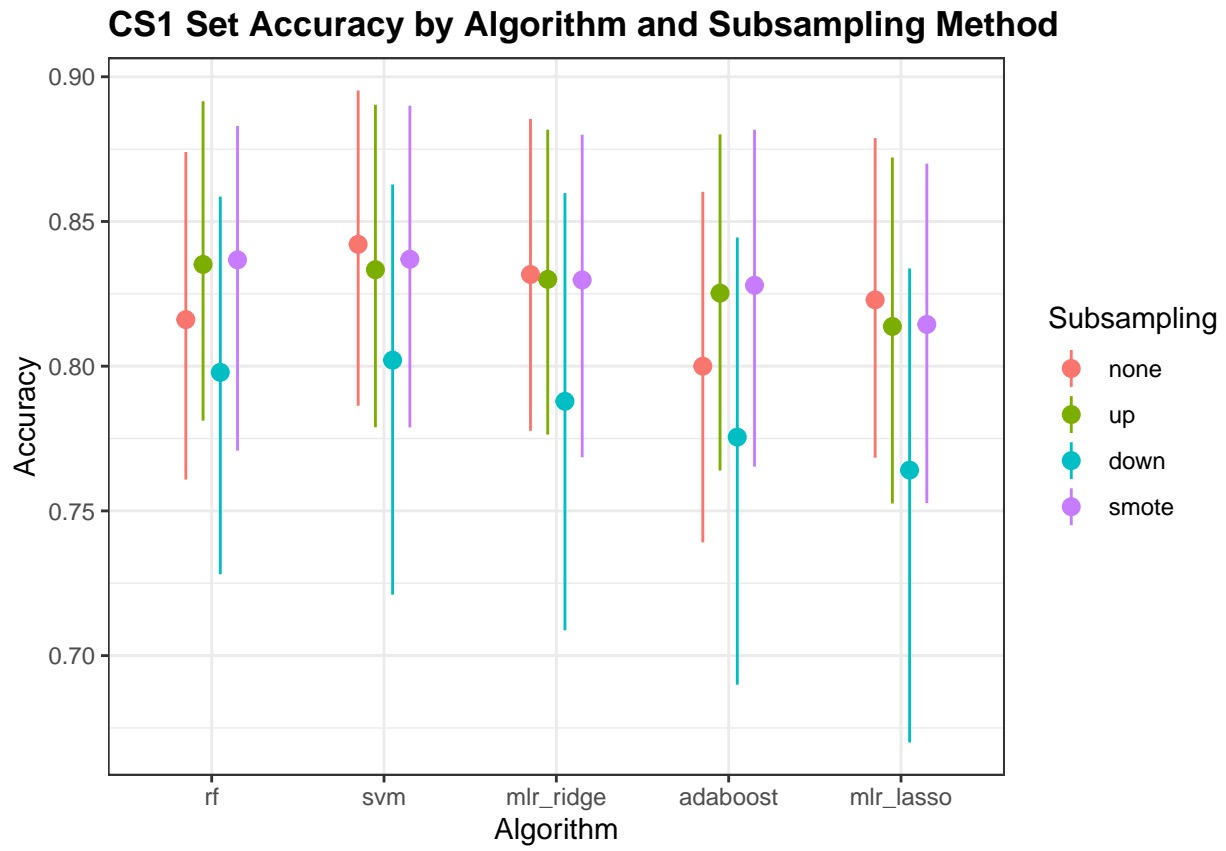


Figure 4.25: CS1 Set Accuracy

Table 4.26: 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 |

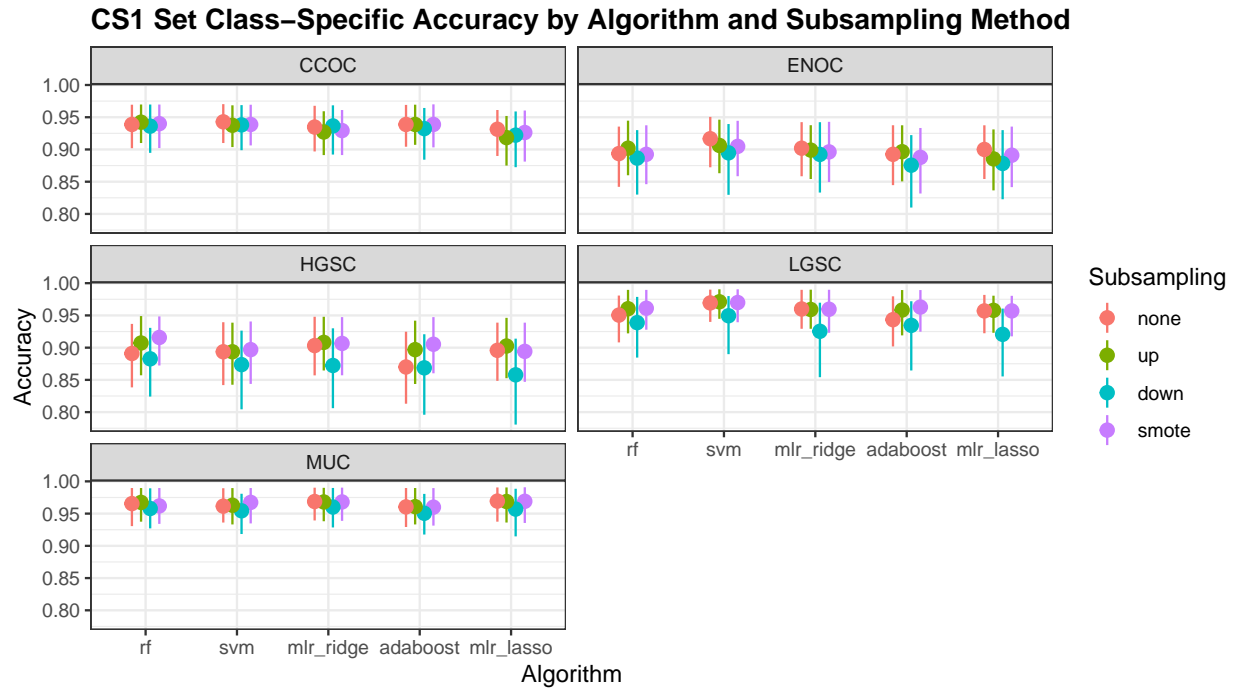


Figure 4.26: CS1 Set Class-Specific Accuracy

Table 4.27: 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

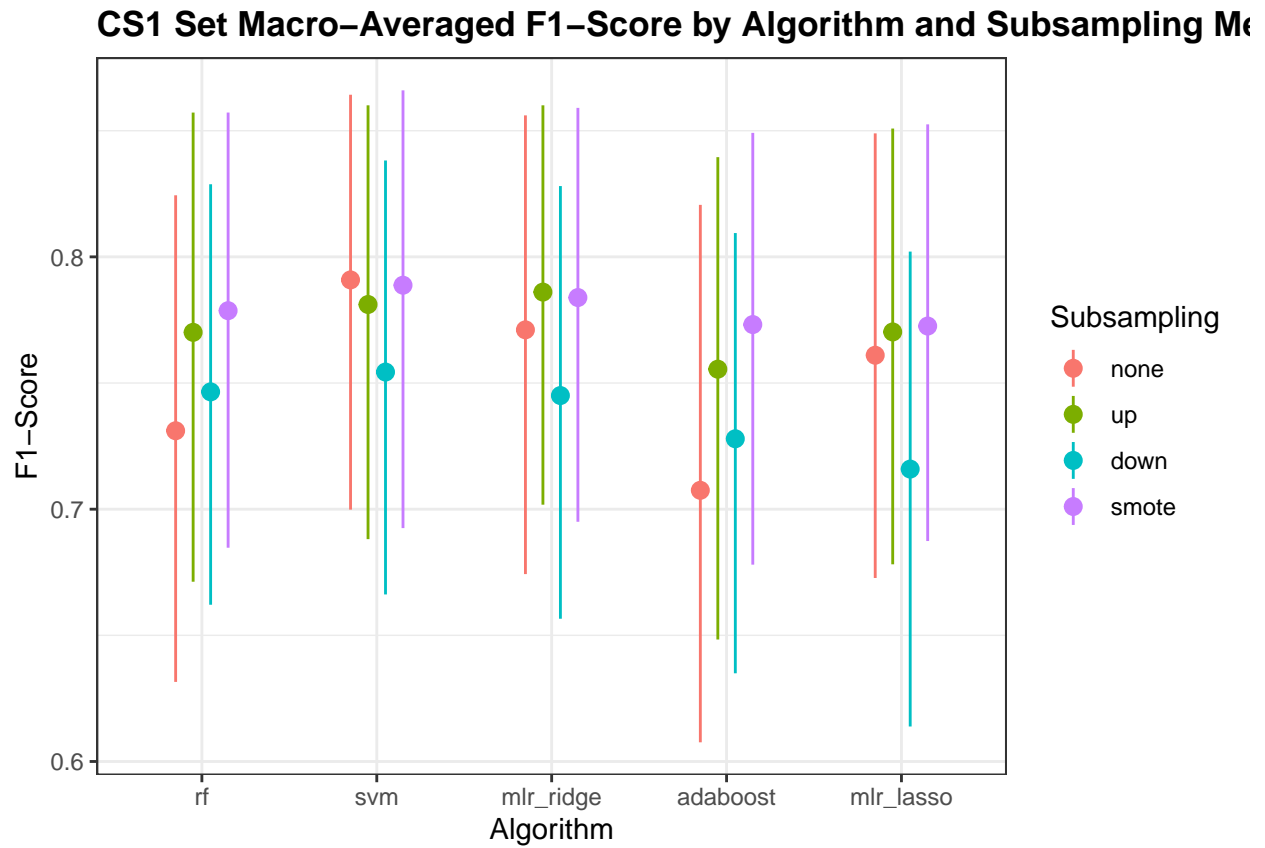


Figure 4.27: CS1 Set F1-Score

Table 4.28: 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 |

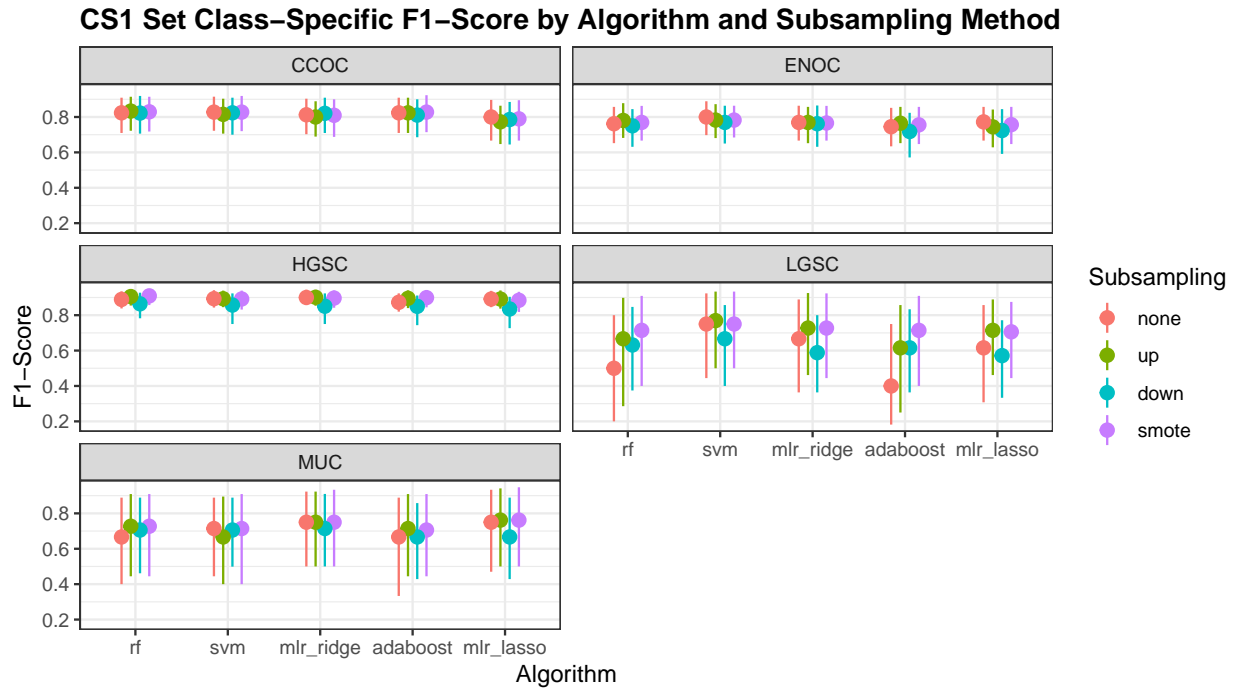


Figure 4.28: CS1 Set Class-Specific F1-Score

Table 4.29: 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

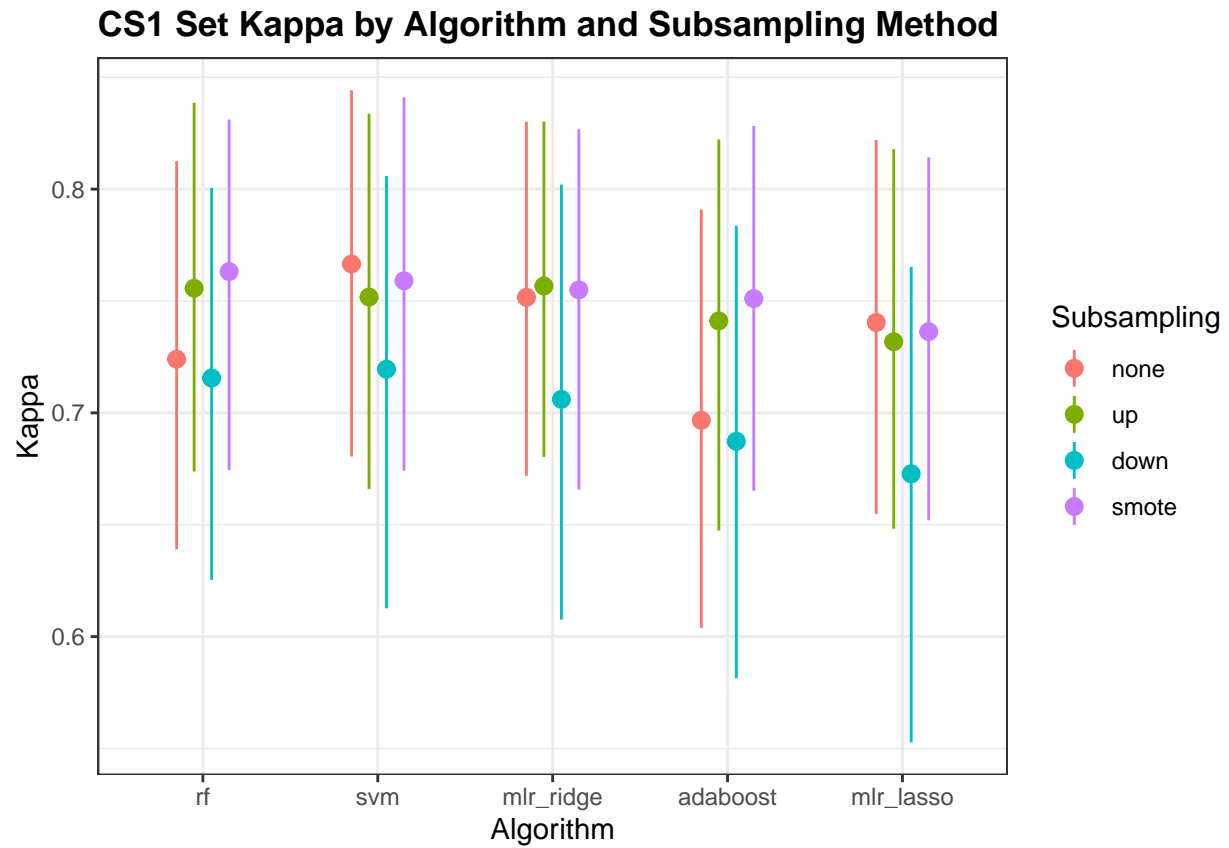


Figure 4.29: CS1 Set Kappa

Table 4.30: 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 |

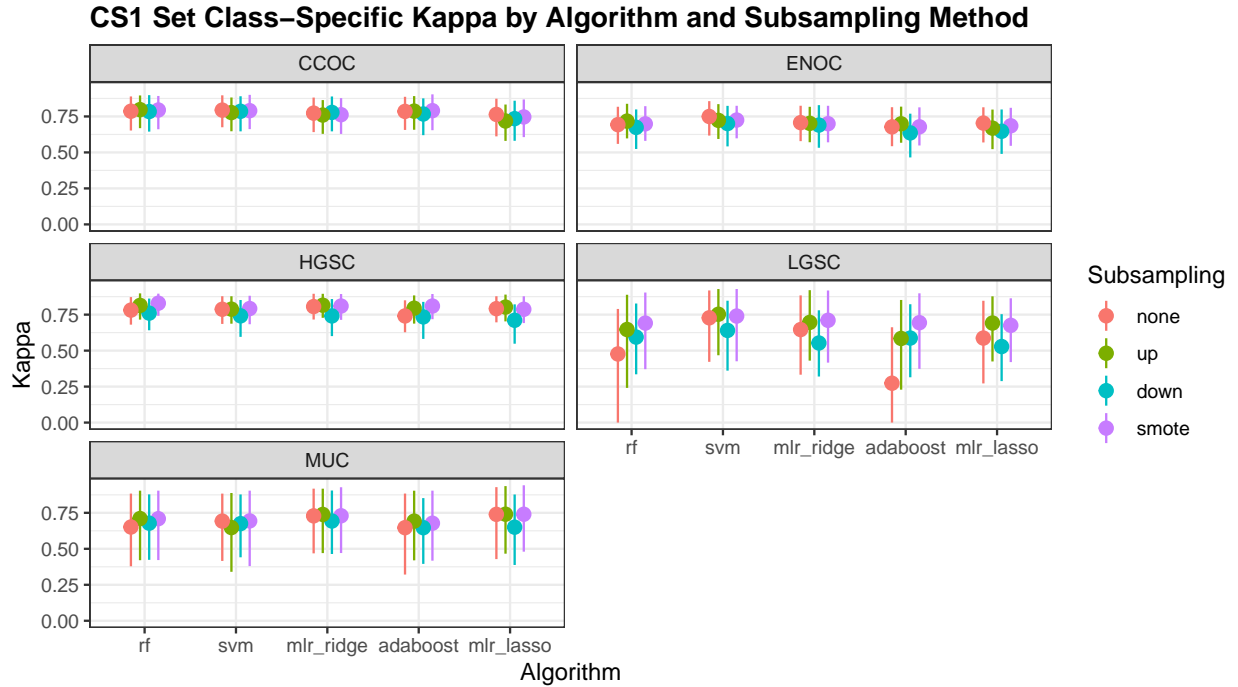


Figure 4.30: CS1 Set Class-Specific Kappa

Table 4.31: 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

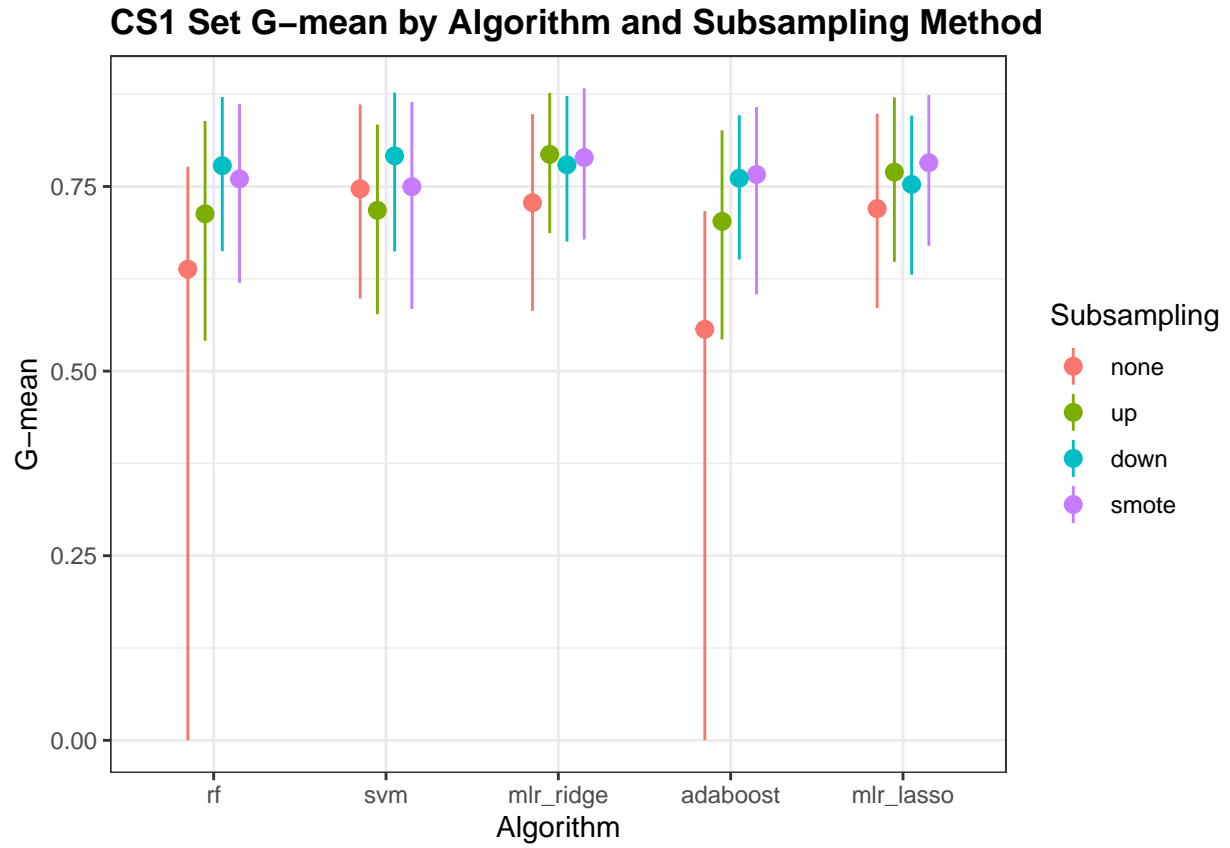


Figure 4.31: CS1 Set G-mean

Table 4.32: 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 |

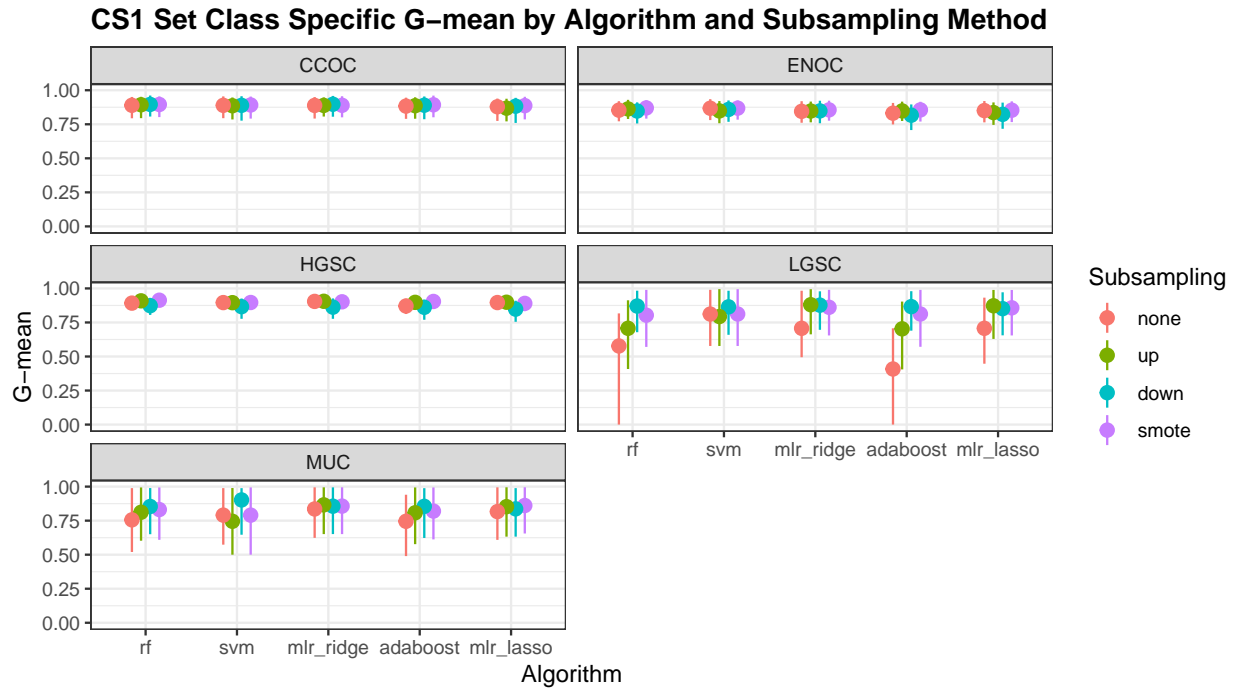


Figure 4.32: CS1 Set Class-Specific G-mean

Table 4.33: 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 CS2 Set

4.4.1 Accuracy

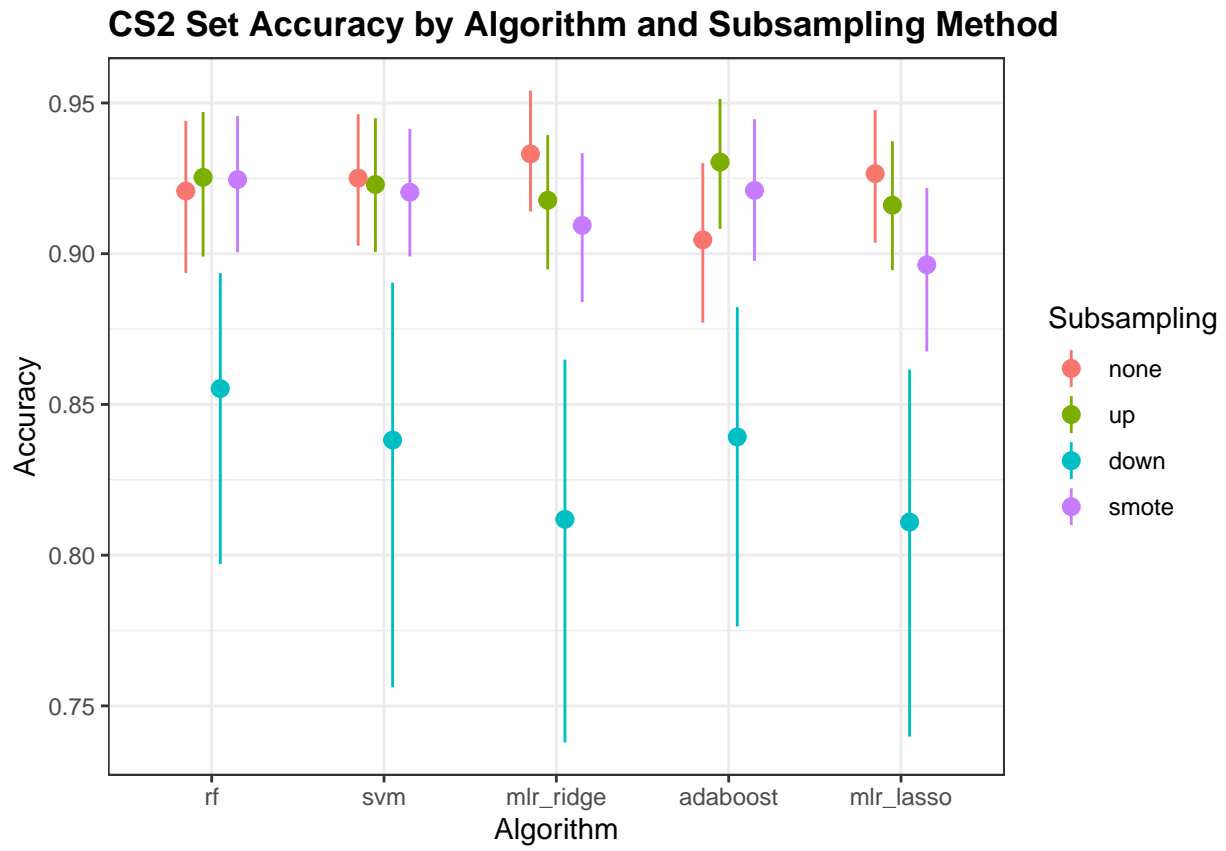


Figure 4.33: CS2 Set Accuracy

Table 4.34: 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 |

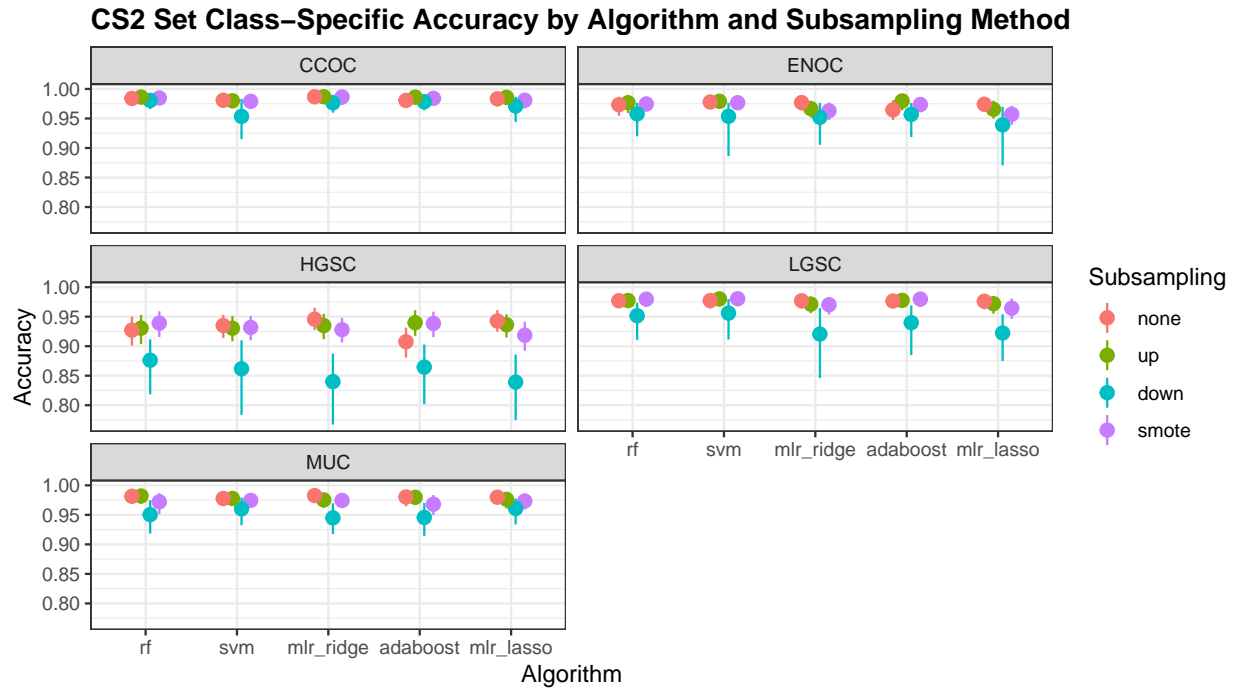


Figure 4.34: CS2 Set Class-Specific Accuracy

Table 4.35: 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

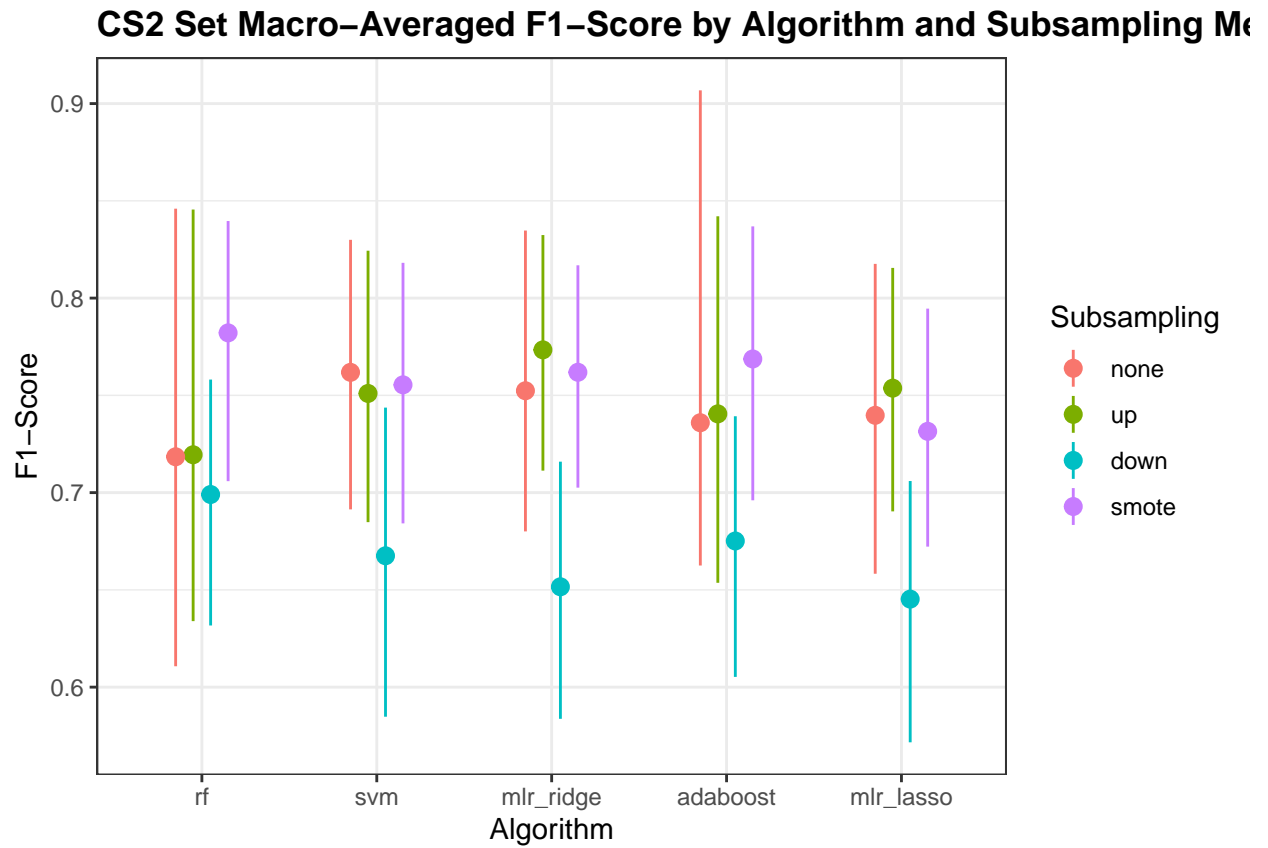


Figure 4.35: CS2 Set F1-Score

Table 4.36: 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 |

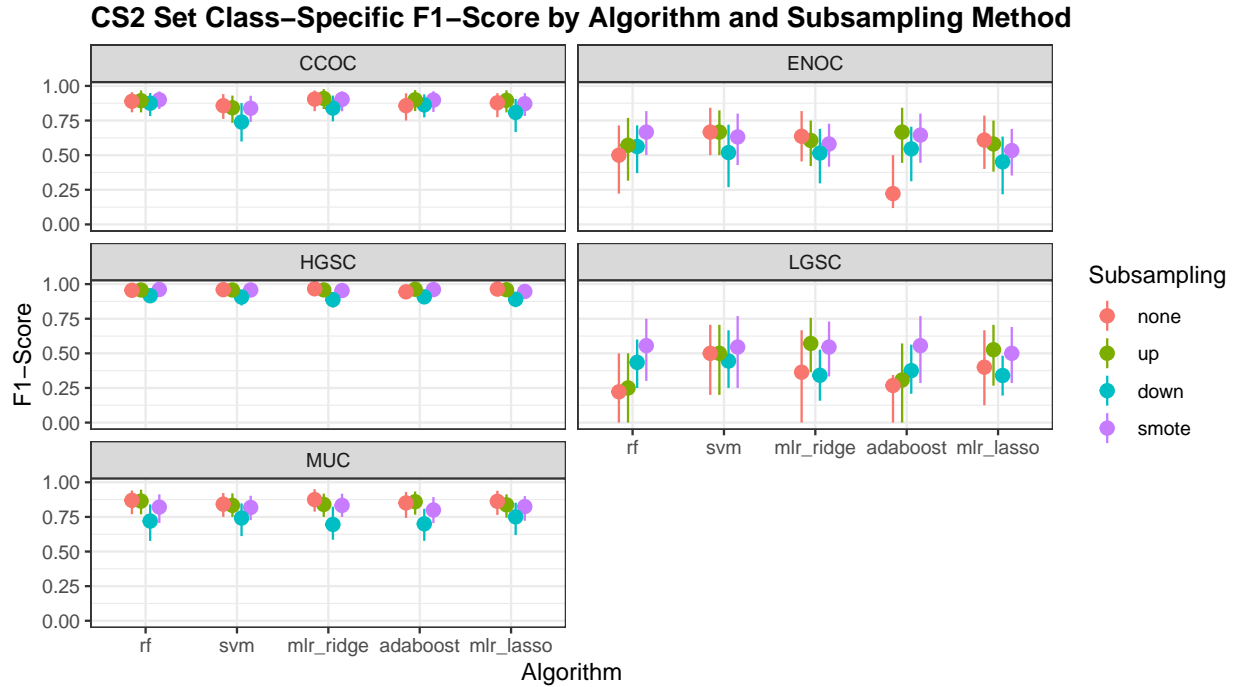


Figure 4.36: CS2 Set Class-Specific F1-Score

Table 4.37: 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

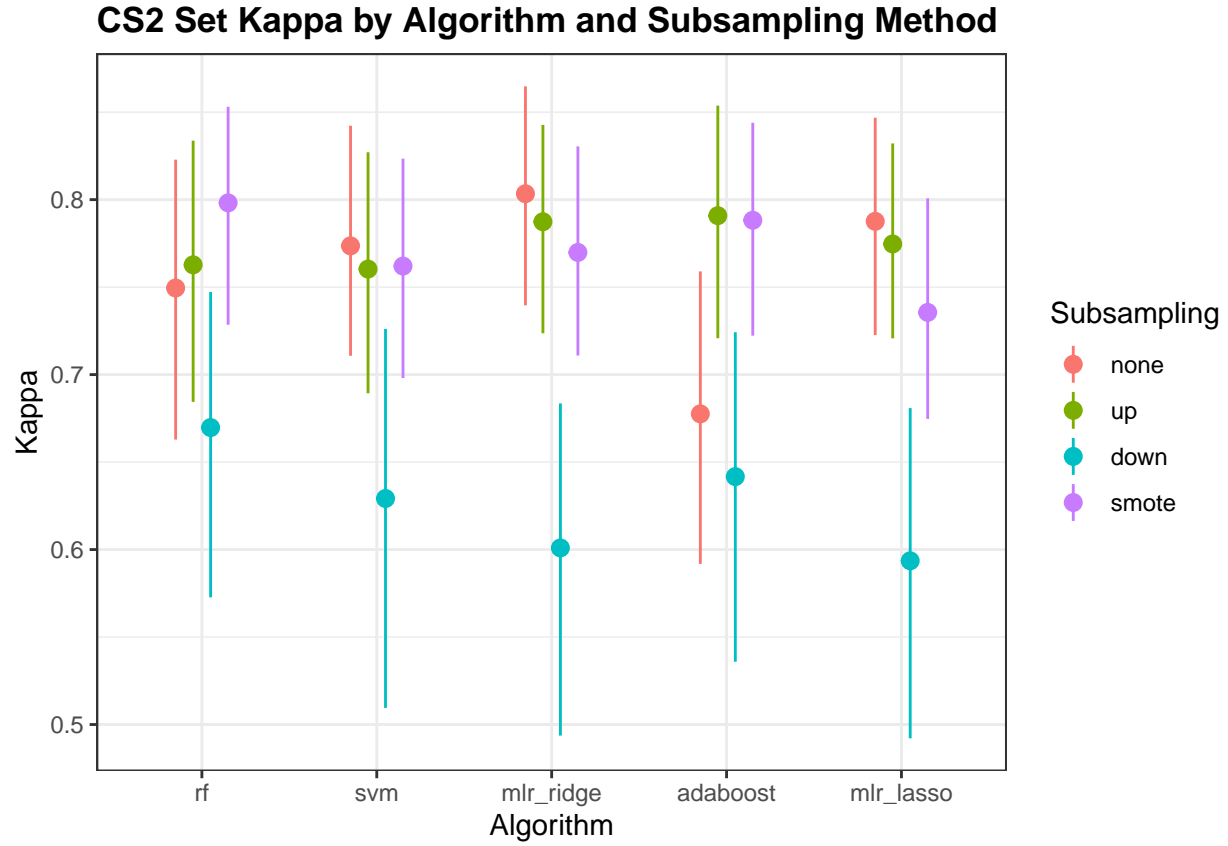


Figure 4.37: CS2 Set Kappa

Table 4.38: 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 |

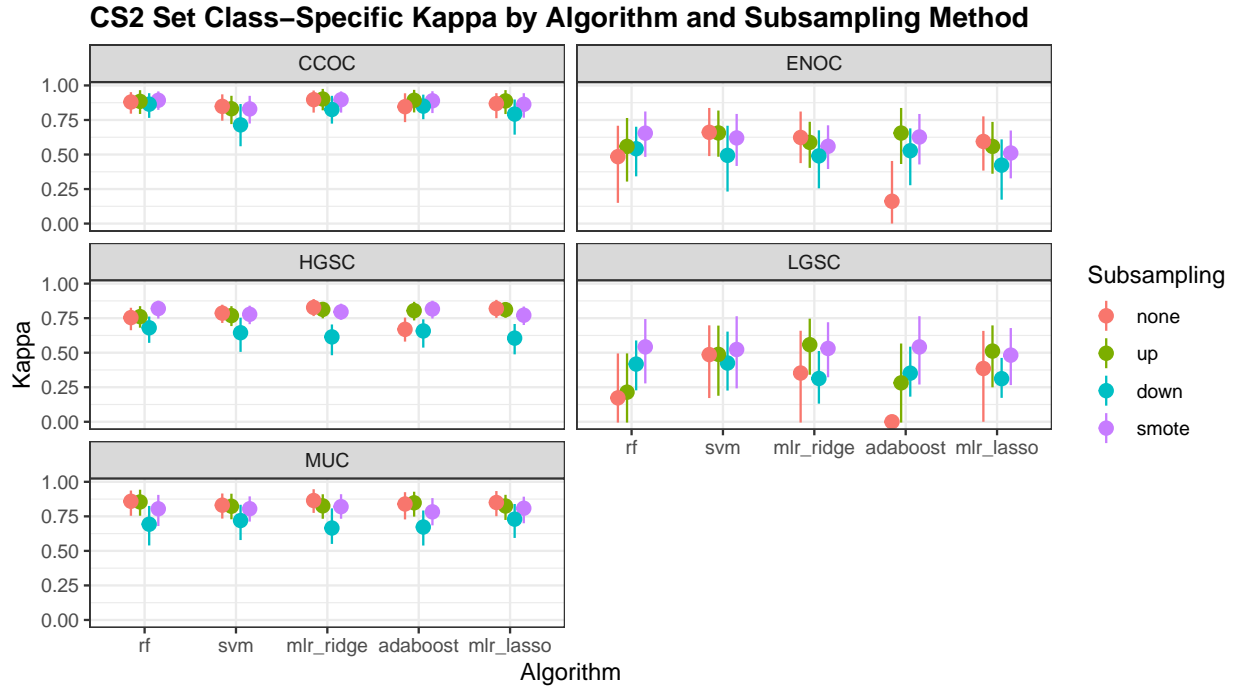


Figure 4.38: CS2 Set Class-Specific Kappa

Table 4.39: 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

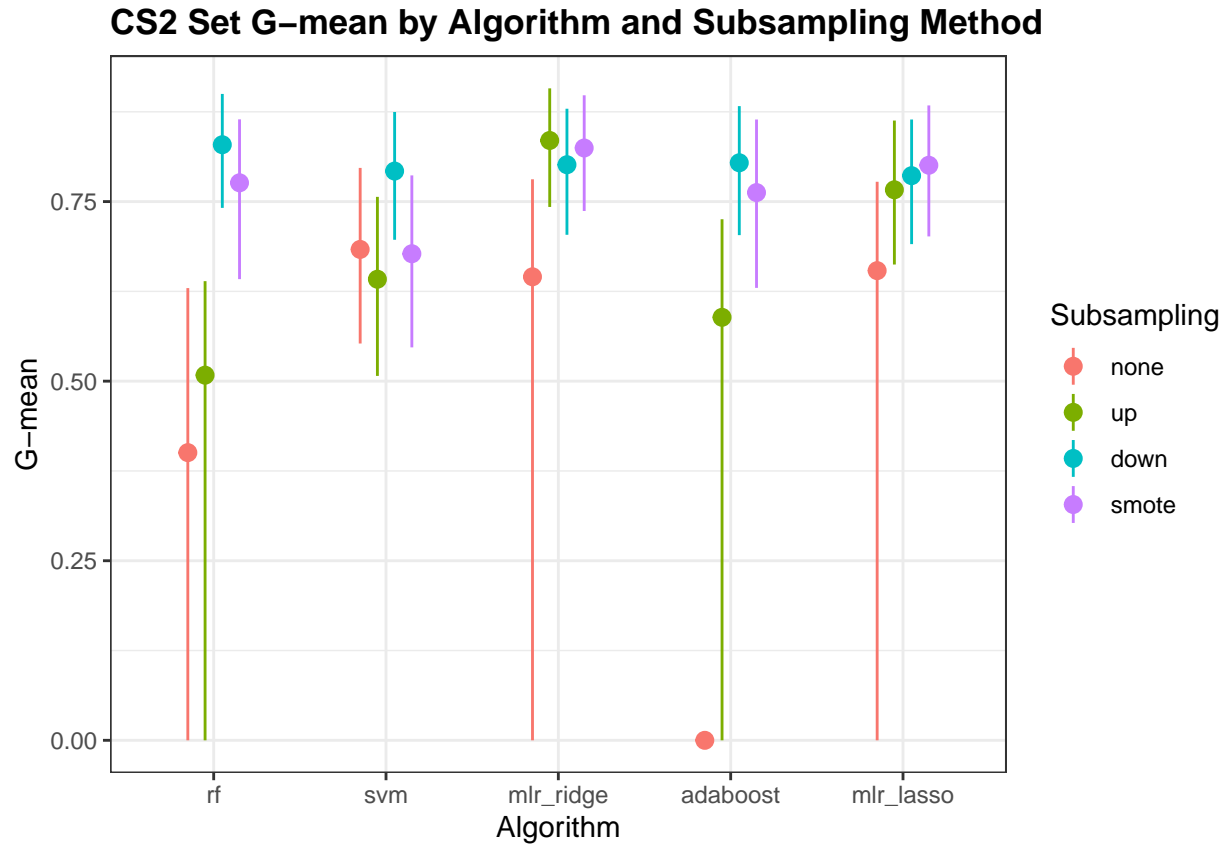


Figure 4.39: CS2 Set G-mean

Table 4.40: 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 |

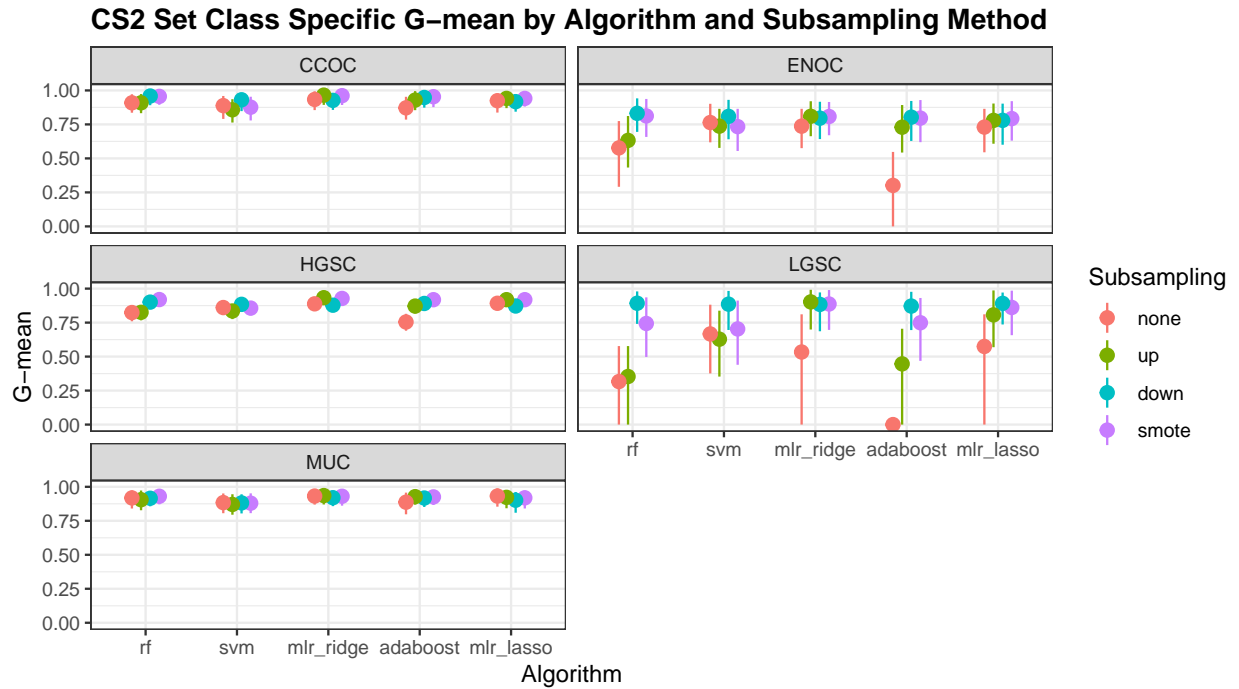


Figure 4.40: CS2 Set Class-Specific G-mean

Table 4.41: 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

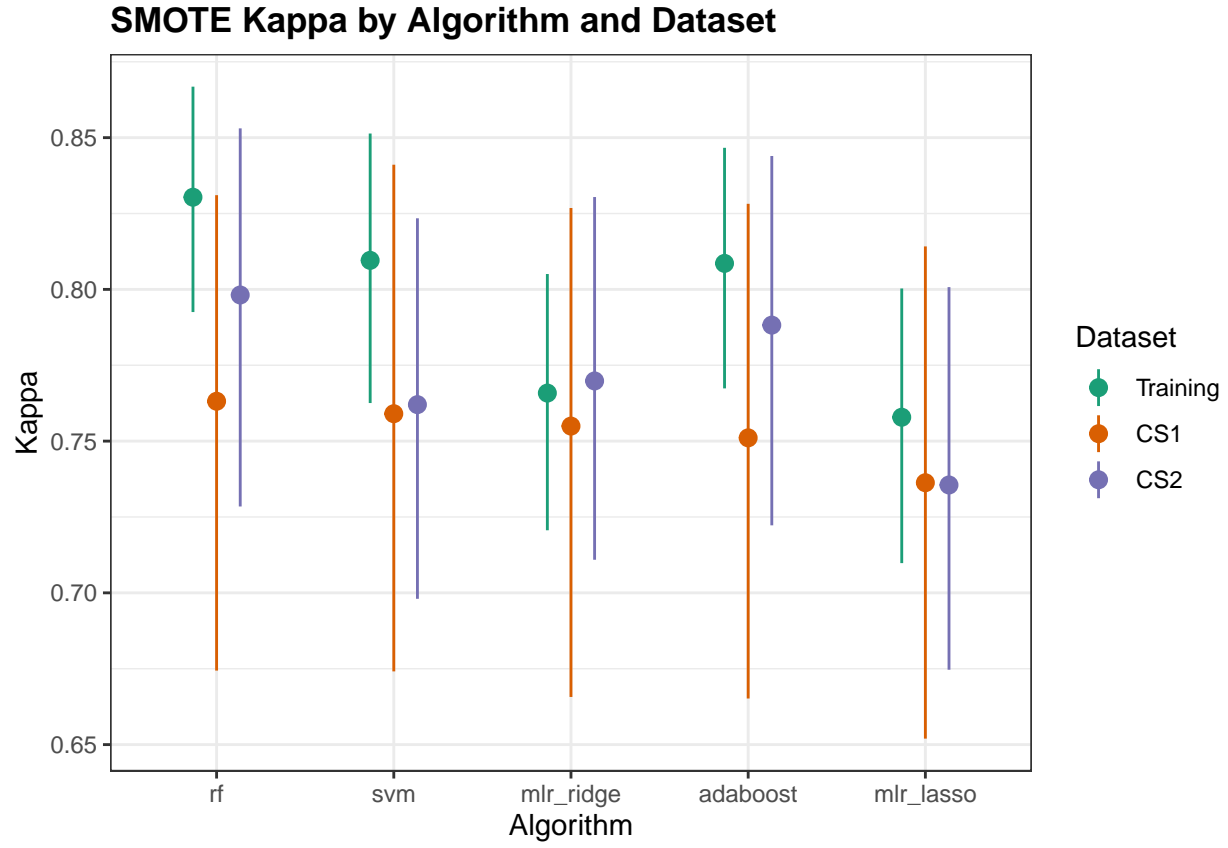


Figure 4.41: SMOTE Kappa by Algorithm and Dataset

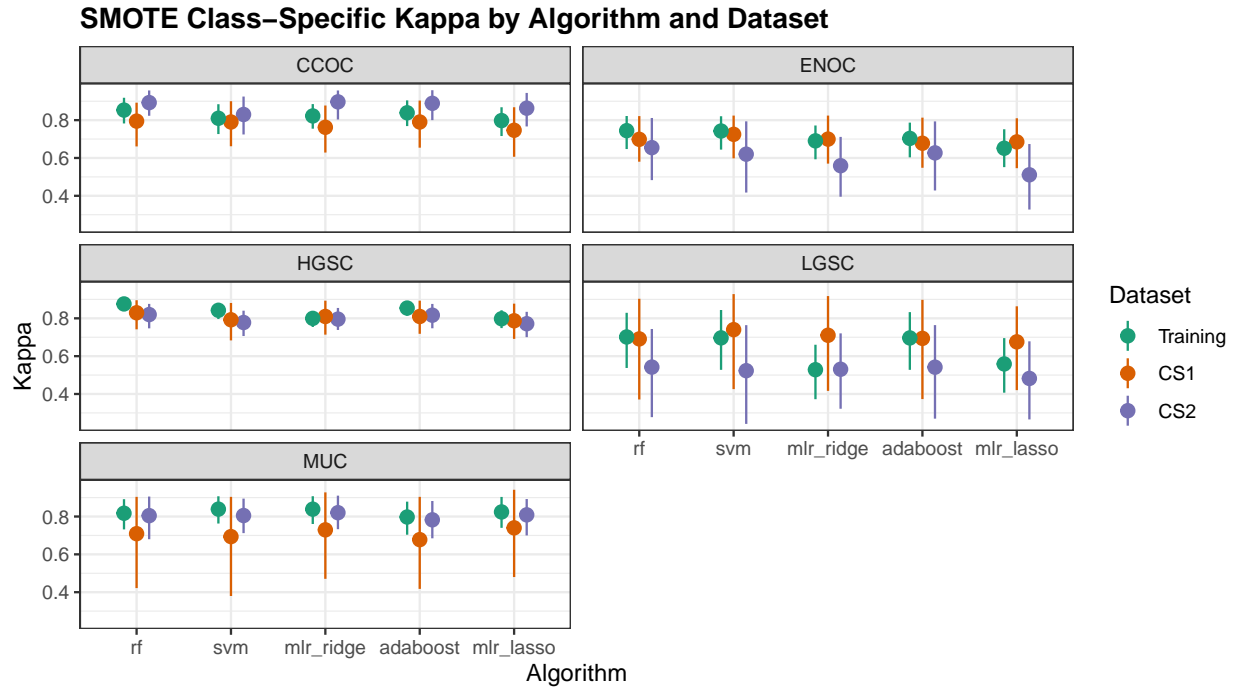


Figure 4.42: SMOTE Class-Specific Kappa by Algorithm and Dataset

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.