Ovarian Cancer Histotypes: Report of Statistical Findings

Derek Chiu

2020-04-09

Contents

List of Tables

List of Figures

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 Aline Talhouk lead the project.

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

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 CS1 to CS3 where we use common samples and genes as reference sets.

2.1.1 Raw Data

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. 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

2.1.4 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.

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 = 9 (Pool 1 = 3, Pool 2 = 3, Pool 3 = 3)
  - 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.

2.2 Classification

We use 6 classification algorithms and 4 subsampling methods across 500 repetitions in the supervised learning framework for CS1 and CS2. The pipeline was run using many SGE batch jobs as a way of parallelization on a CentOS 5 server. Implementations of the techniques below were called from the splendid package.

- Classifiers:
  - Random Forest
  - Adaboost
  - XGBoost
  - LDA
  - SVM
  - K-Nearest Neighbours
- Subsampling:
  - None
  - Down-sampling
  - Up-sampling
  - SMOTE

3. Validation

3.1 Concordance

First we validate the CS2 normalization process by looking at the distribution of CS3 non-normalized samples with:

- CS2 non-normalized
- CS2 pools-normalized
- CS2 reference-normalized

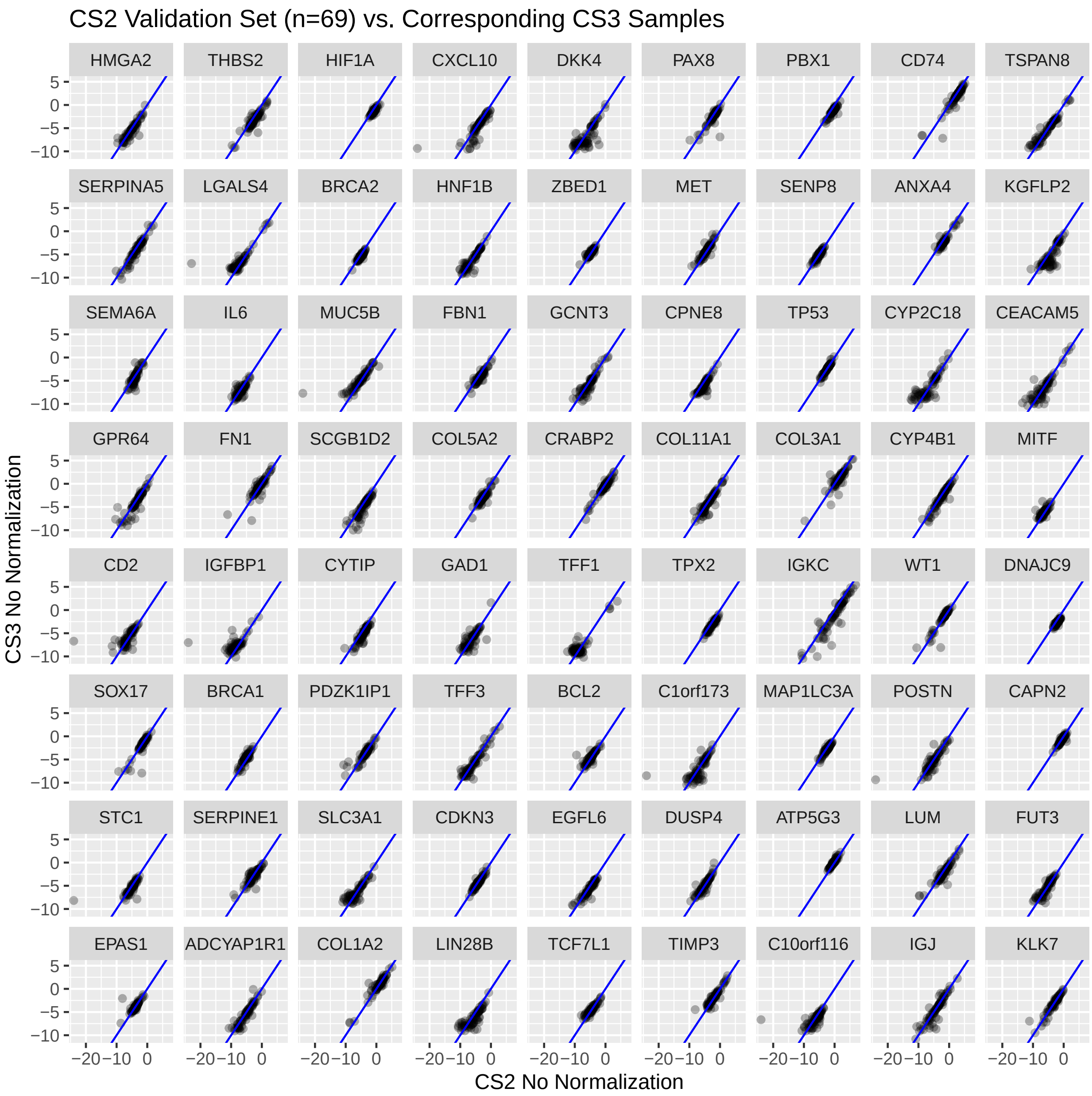

Figure 3.1: Gene Expression CS2 No Normalization vs. CS3

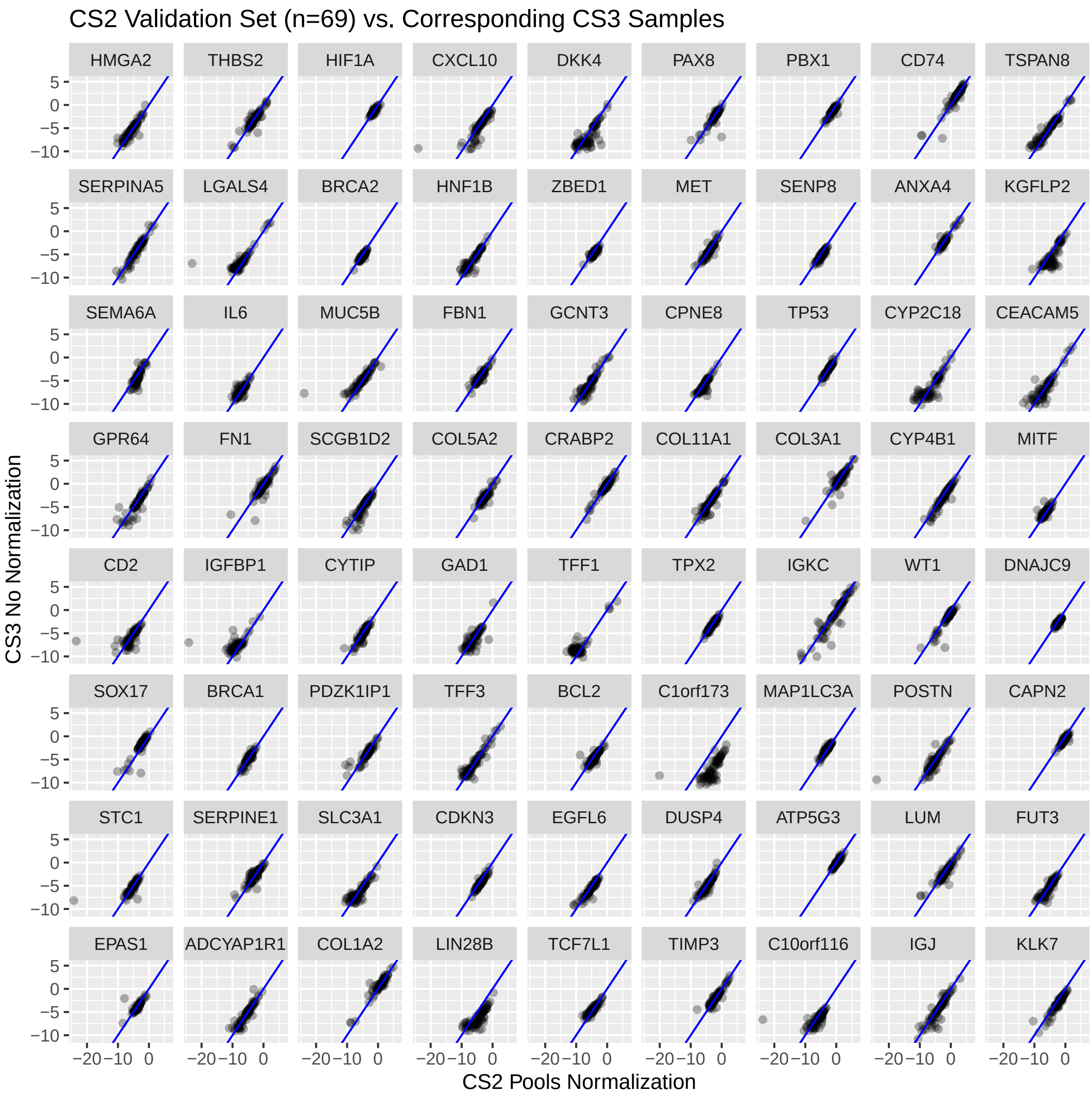

Figure 3.2: Gene Expression CS2 Pools Normalization vs. CS3

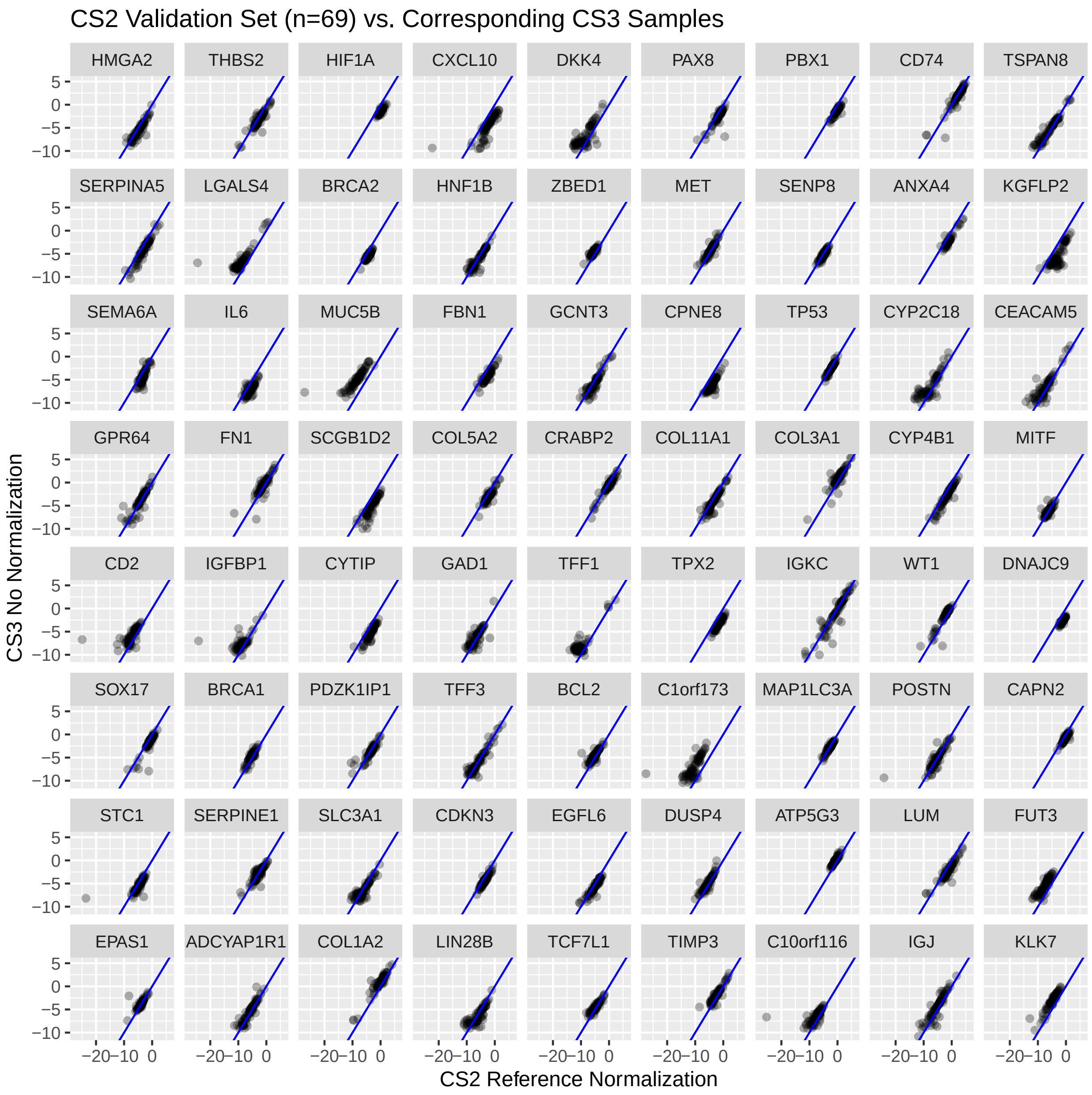

Figure 3.3: Gene Expression CS2 Reference Normalization vs. CS3

Table 3.1: All CodeSet Histotype Groups

hist_gr	CS1	CS2	CS3
HGSC	169	757	2453
non-HGSC	196	377	677

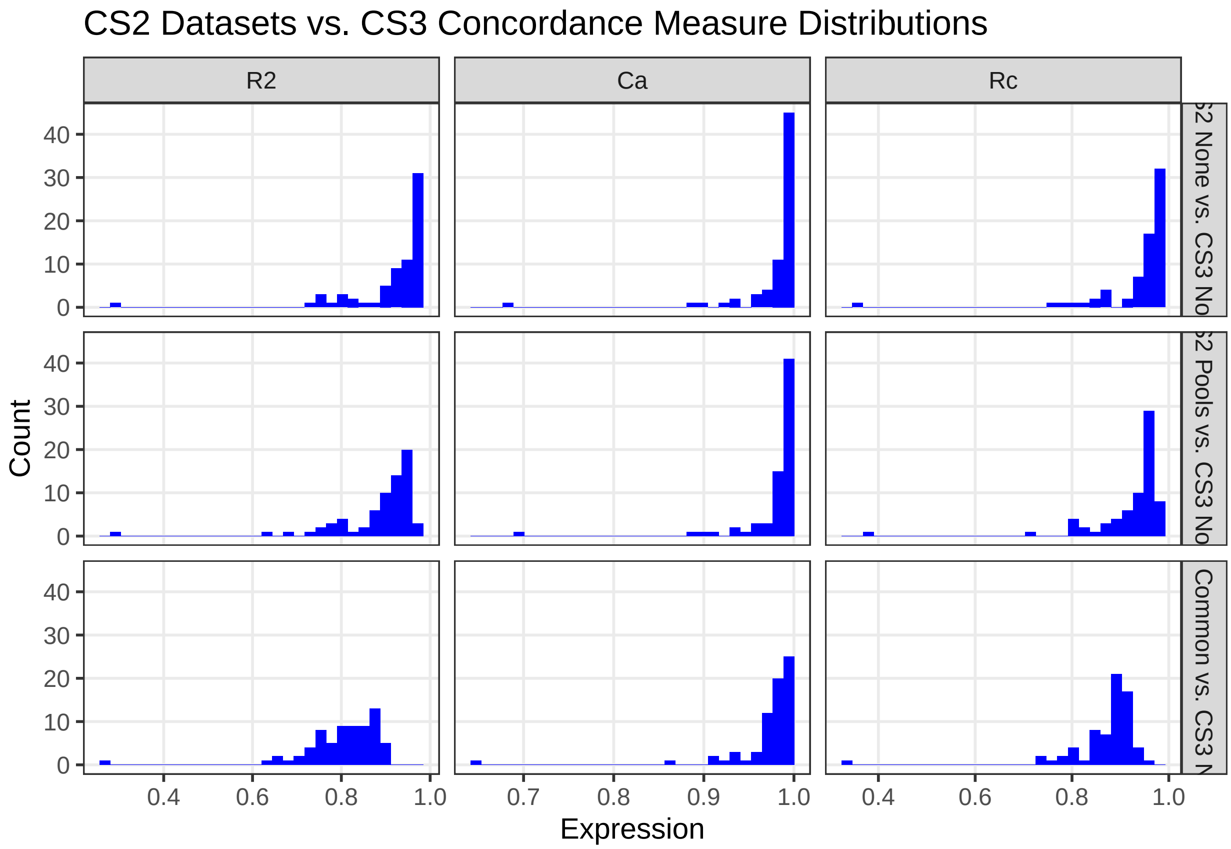

Figure 3.4: Concordance Histograms

3.2 Full Data Distributions

The histotype distributions on the full data are shown below.

3.3 Training Set Distributions

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

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
UNK	0	4	0

Table 3.3: 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 3.4: 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
CS2	UNK	4

Table 3.5: 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 3.6: CS1 Training Set Histotypes

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

Table 3.7: CS2 Training Set Histotypes

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

4. Results

Here we show internal validation summaries for both CS1 and CS2. The accuracy and F1-scores are the measures of interest. Algorithms are sorted by descending value. The point ranges show the median, 5th and 95th percentiles, coloured by subsampling methods.

4.1 CS1

4.1.1 Accuracy

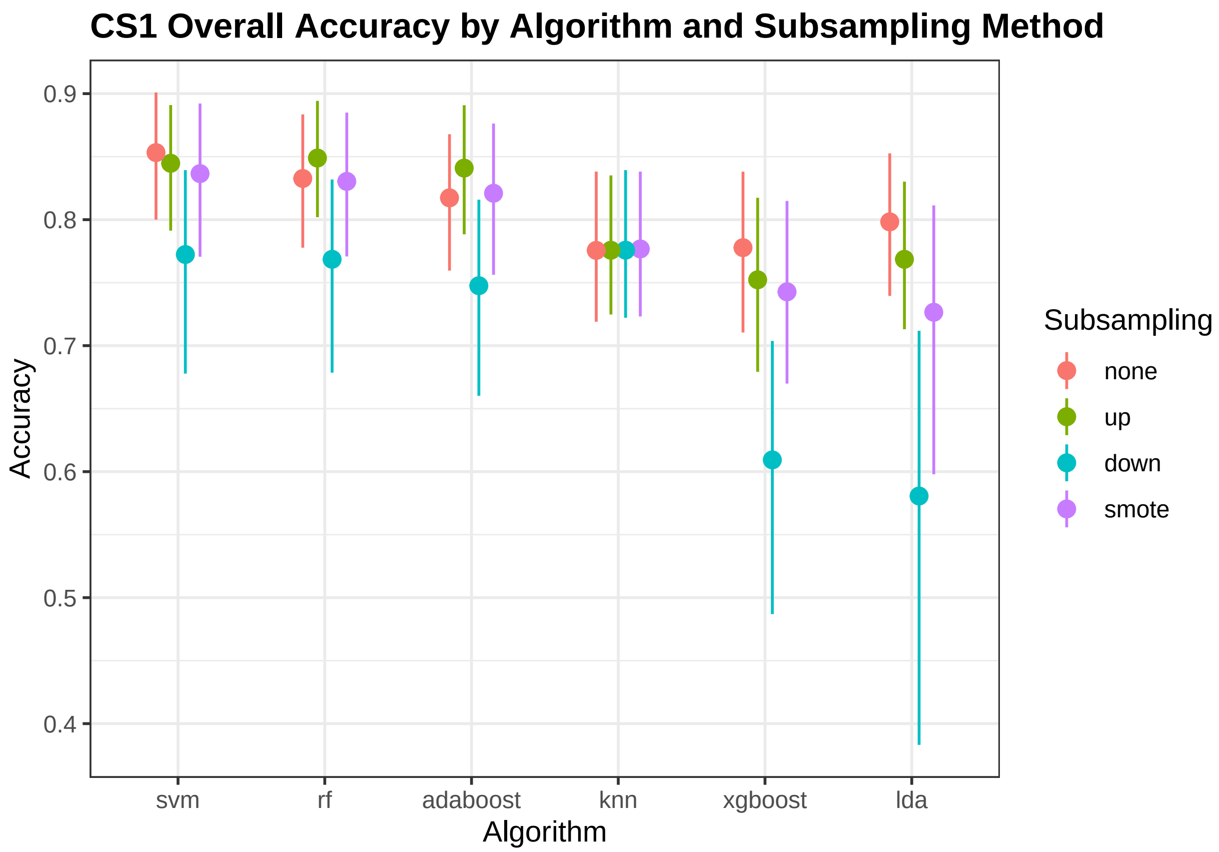

Figure 4.1: CS1 Accuracy

4.1.2 F1-Score

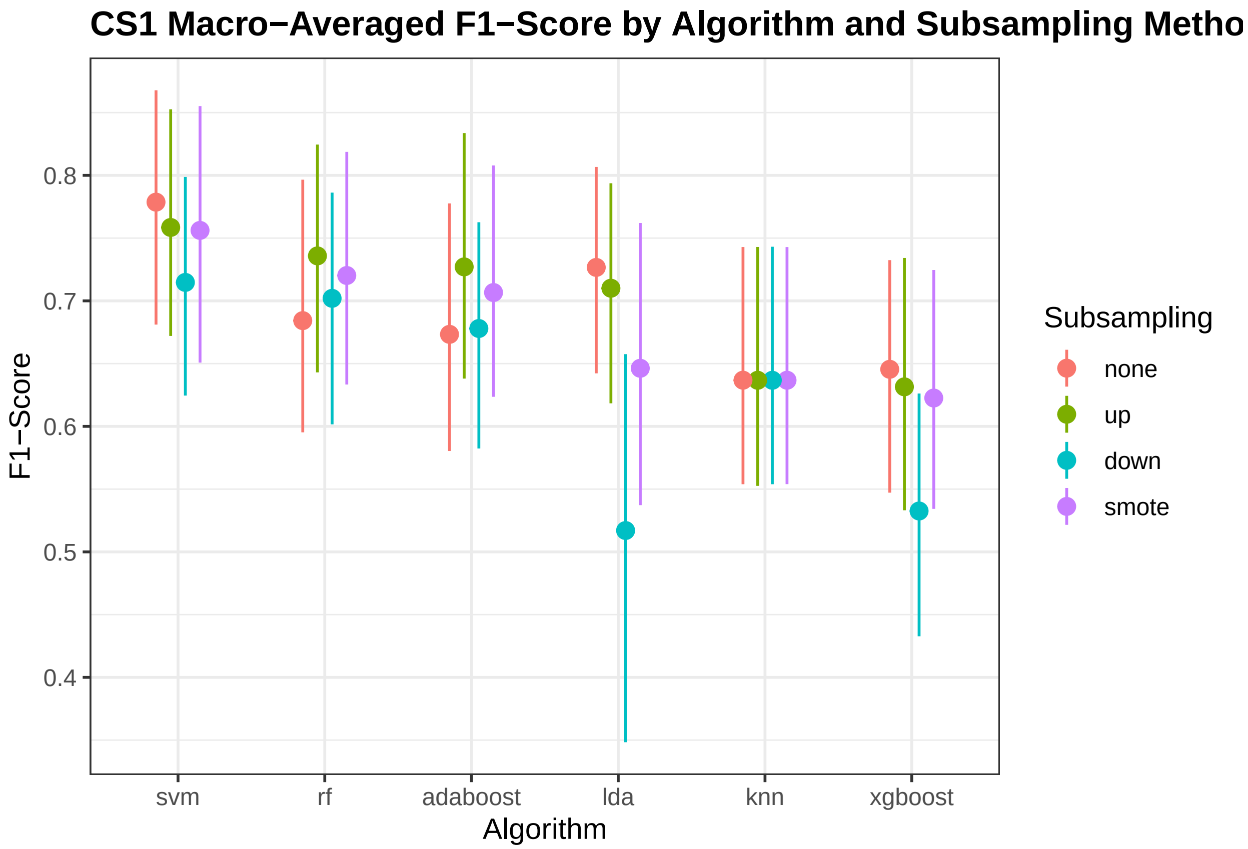

Figure 4.2: CS1 F1-Score

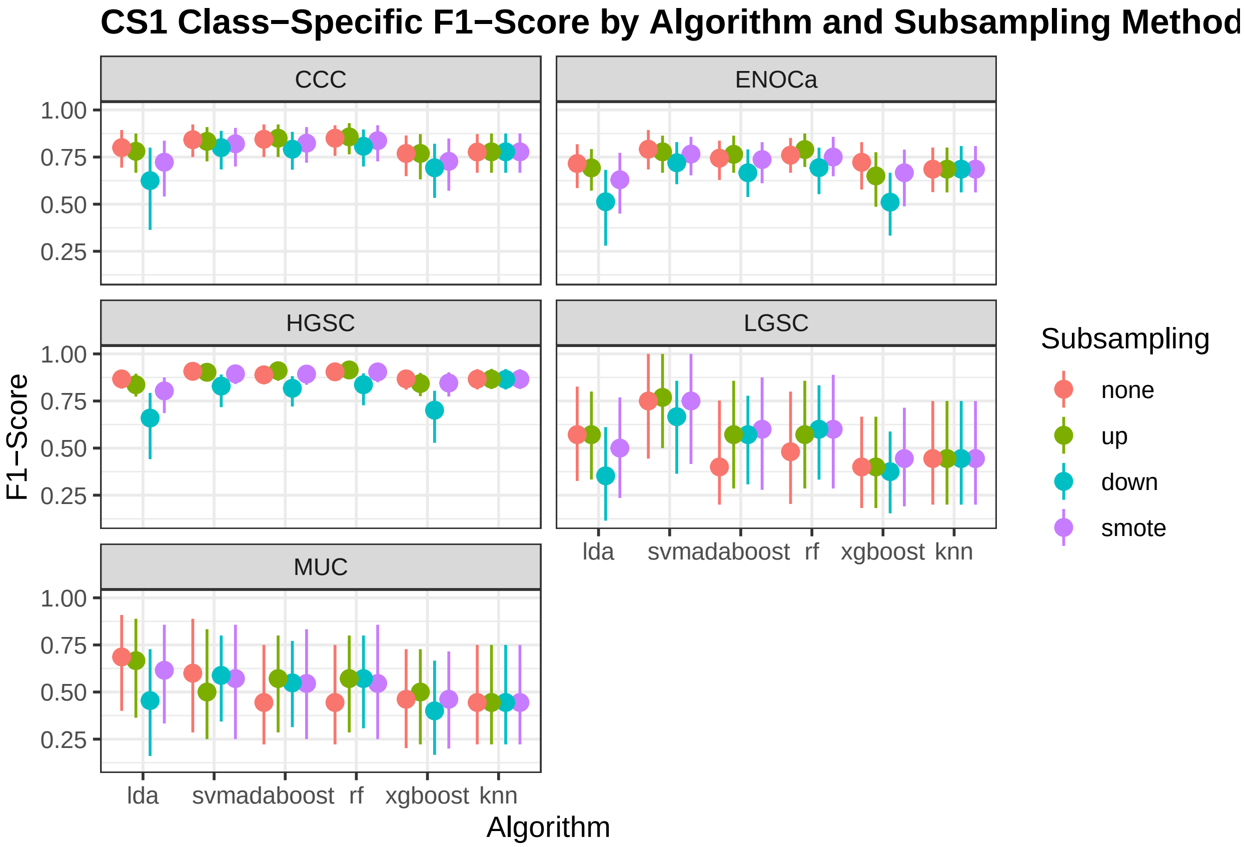

Figure 4.3: CS1 Class-Specific F1-Score

4.2 CS2

4.2.1 Accuracy

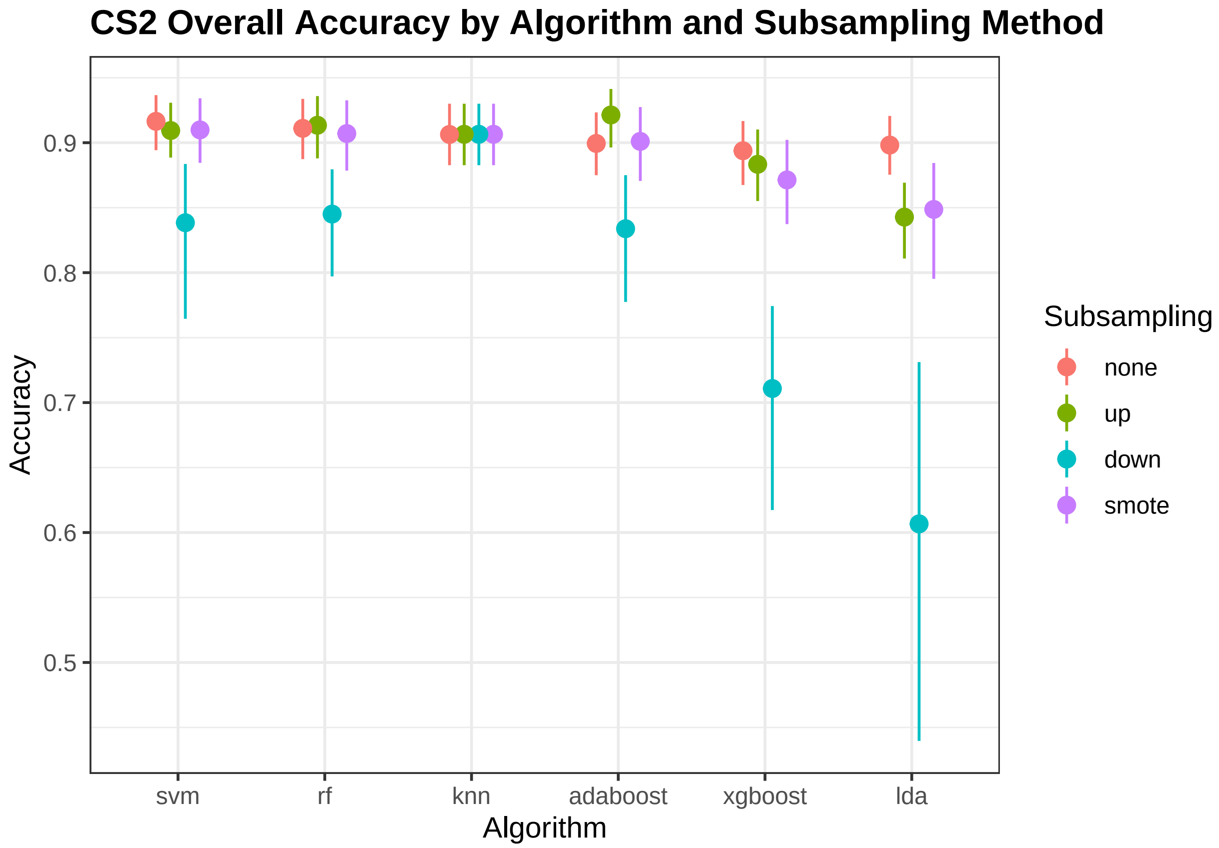

Figure 4.4: CS2 Accuracy

4.2.2 F1-Score

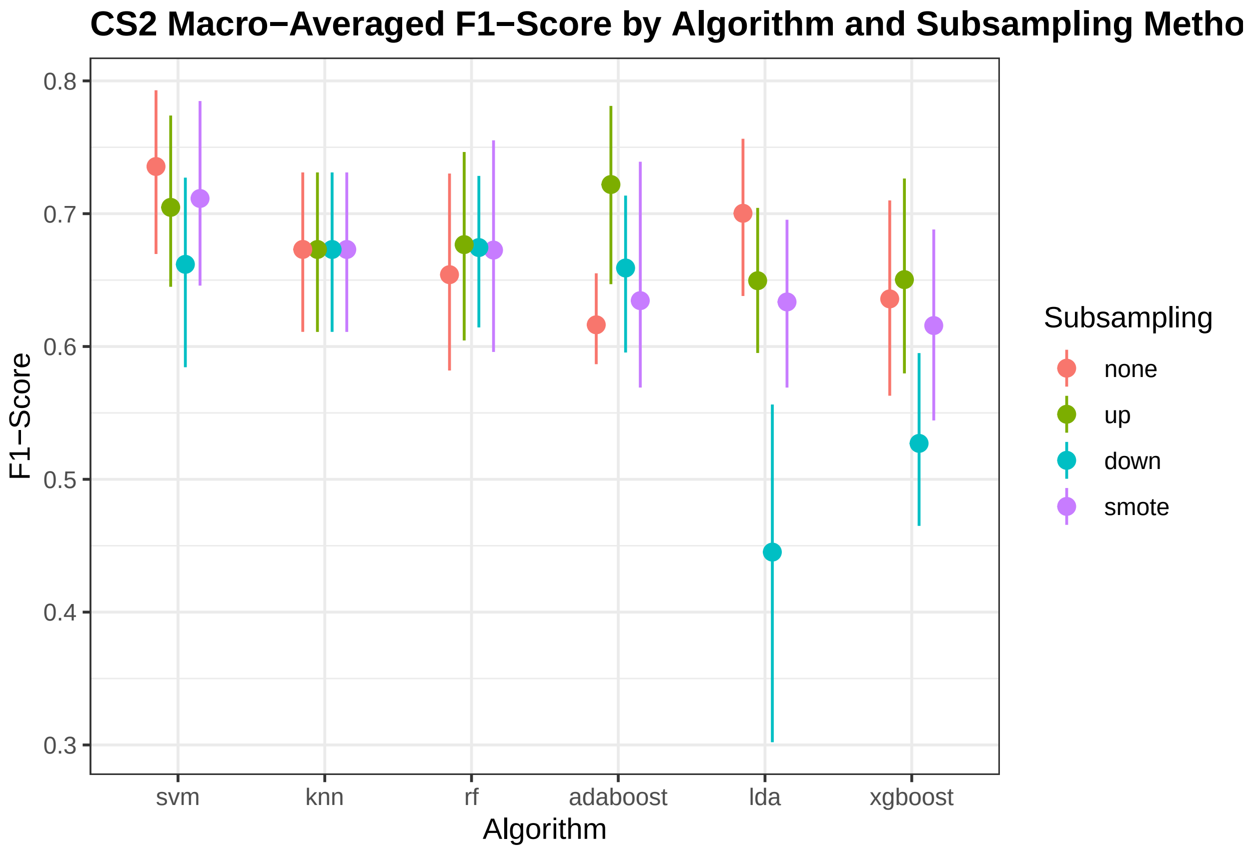

Figure 4.5: CS2 F1-Score

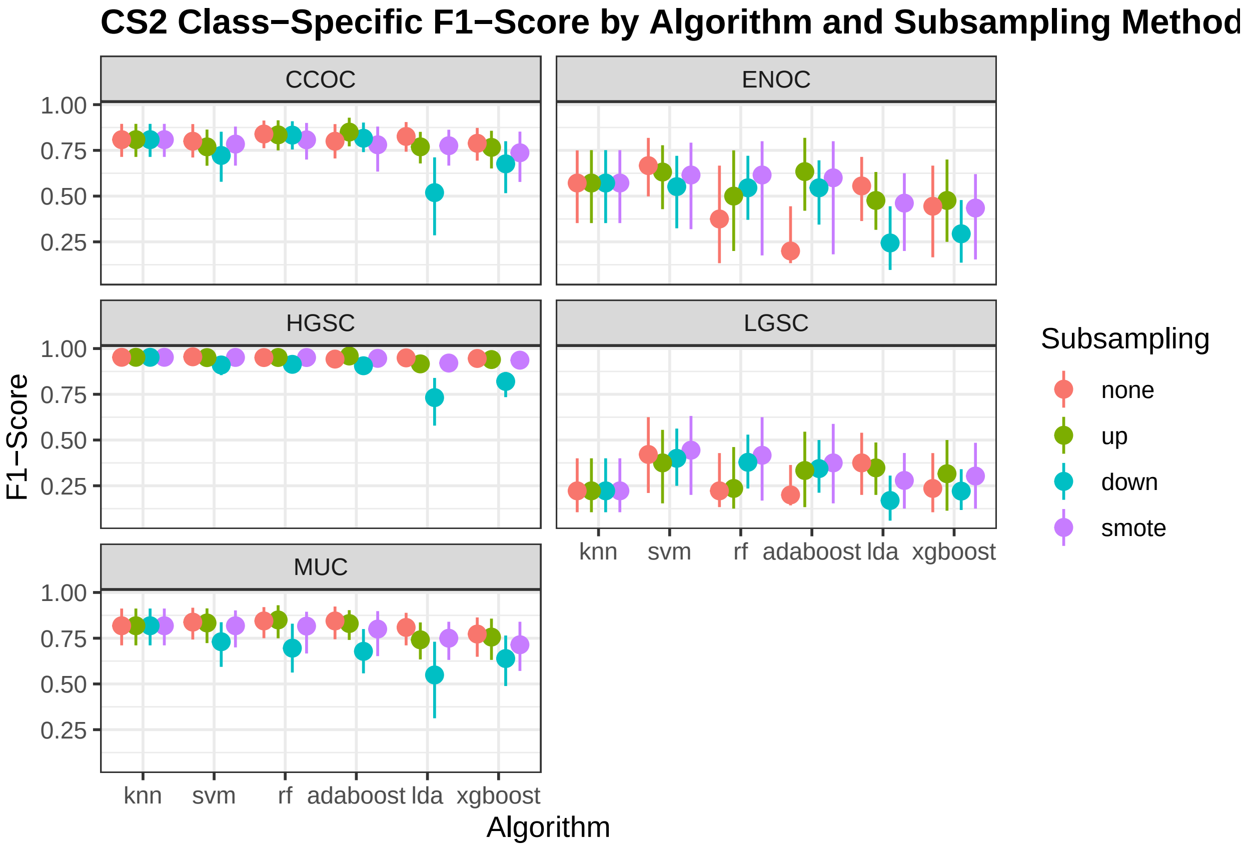

Figure 4.6: CS2 Class-Specific F1-Score