Ovarian Cancer Histotypes: Report of Statistical Findings

Derek Chiu

June 5, 2026

Table of contents

List of Figures

List of Tables

Preface

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

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

1 Introduction

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

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

2 Methods

2.1 Pre-Processing

2.1.1 Case Selection

Prior to pre-processing, samples were split into a training, a confirmation, and a validation set.

- Training
  - CS1: OOU, OOUE, VOA, MAYO, MTL
  - CS2: OOU, OOUE, VOA, MAYO, OVAR3, OVAR11, JAPAN, MTL, POOL-CTRL
  - CS3: OOU, OOUE, VOA, POOL-1, POOL-2, POOL-3
- Confirmation:
  - CS3: TNCO
- Validation:
  - CS3: DOVE4

2.1.2 Quality Control

Before normalization, we calculated several quality control measures and excluded samples that failed to achieve sample quality in one or more of these measures.

- **Linearity of positive control genes**: If the R-squared from a linear model of positive controls and their concentrations is less than 0.95 or missing, then the sample is flagged.
- **Imaging quality**: The sample is flagged if the field of view percentage is less than 75%.
- **Positive Control flag**: We consider the two smallest positive controls at concentrations 0.5 and 1. If these two controls are less than the lower limit of detection (defined as two standard deviations below the mean of the negative control expression), or if the mean negative control expression is 0, the sample is flagged.
- **The signal-to-noise ratio or percent of genes detected**: These two measures are defined as the ratio of the average housekeeping gene expression over the upper limit of detection, defined as two standard deviations above the mean of the negative control expression (or 0 if this limit is less than 0.001), and the proportion of endogenous genes with expression greater than the upper limit of detection. These measures are flagged if they are below a pre-specified threshold, which is determined visually by considering their bivariate distribution in a scatterplot. In this case, we used 100 for the SNR threshold and 50% for the threshold for genes detected. Note: these thresholds were determined by examining the relationship in Section 3.3.2.

2.1.3 Housekeeping Genes Normalization

The full training set (n=1257) comprised of data from three CodeSets (CS) 1, 2, and 3. Data normalization removes technical variation from high-throughput platforms to improve the validity of comparative analyses.

Each CodeSet was first normalized to housekeeping genes: *ACTB*, *RPL19*, *POLR1B*, *SDHA*, and *PGK1*. Housekeeping genes encode proteins responsible for basic cell function and have consistent expression in all cells. All expression values were log2 transformed. Normalization to housekeeping genes corrects the viable RNA from each sample. This is achieved by subtracting the average log (base 2)-transformed expression of the housekeeping genes from the log (base 2)-transformed expression of each gene:

$$log_2(\text{endogenous gene expression}) - \text{average}(log_2(\text{housekeeping gene expresssion})) = \text{relative expression} \tag{2.1}$$

2.1.4 Between CodeSet and Site Normalization

To normalize between CodeSets, we randomly selected five specimens, one from each histotype, among specimens repeated in all three CodeSets. This formed the reference set (Random 1). We selected only one sample from each histotype to use as few samples as possible for normalization and retain the rest for analysis.

A reference-based approach (Talhouk et al. (2016)) was used to normalize CS1 to CS3 and CS2 to CS3 across their common genes:

$$\begin{aligned} \text{X-Norm}_{\text{CS1}} &= X_{\text{CS1}} + \bar{R}_{\text{CS3}} - \bar{R}_{\text{CS1}} \\ \text{X-Norm}_{\text{CS2}} &= X_{\text{CS2}} + \bar{R}_{\text{CS3}} - \bar{R}_{\text{CS2}} \end{aligned} \tag{2.2}$$

Samples in CS3 were processed at three different locations; we also had to normalize for "site" in this CodeSet. Finally, the CS3 expression samples were included in the training set without further normalization:

$$\begin{aligned} \text{X-Norm}_{\text{CS3-USC}} &= X_{\text{CS3-USC}} + \bar{R}_{\text{CS3-VAN}} - \bar{R}_{\text{CS3-USC}} \\ \text{X-Norm}_{\text{CS3-AOC}} &= X_{\text{CS3-AOC}} + \bar{R}_{\text{CS3-VAN}} - \bar{R}_{\text{CS3-AOC}} \end{aligned} \tag{2.3}$$

The CS3 expression samples were included in the training set without further normalization. CS3 cross-site replicates were all chosen from Vancouver, thus we did not retain any samples from USC or AOC. Finally, the initial training set is assembled by combining the normalized datasets for CS1 and CS2 along with the CS3 expression not used in normalization:

$$\begin{aligned} \text{Training Set} &= \text{X-Norm}_{\text{CS1}} + \text{X-Norm}_{\text{CS2}} + \text{X}_{\text{CS3-VAN}} + \text{X-Norm}_{\text{CS3-USC}} + \text{X-Norm}_{\text{CS3-AOC}} \\ &= \text{X-Norm}_{\text{CS1}} + \text{X-Norm}_{\text{CS2}} + \text{X}_{\text{CS3}} \end{aligned} \tag{2.4}$$

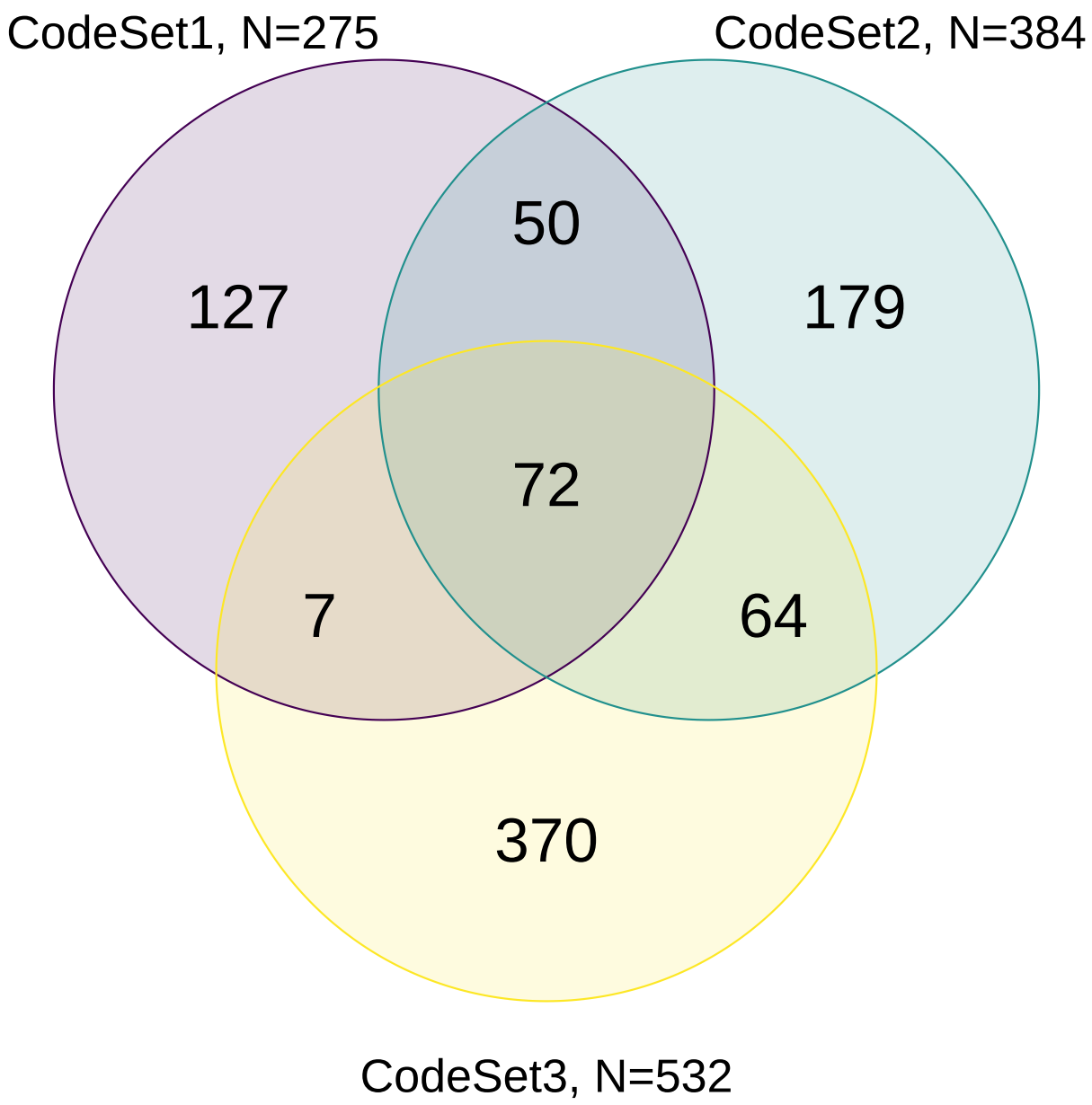


Figure 2.1: Venn diagram of common and unique gene targets covered by each CodeSet

2.1.5 Final Processing

We map ovarian histotypes to all remaining samples and keep the major histotypes for building the predictive model: high-grade serous ovarian carcinoma (HGSOC), clear cell ovarian carcinoma (CCOC), endometrioid ovarian carcinoma (ENOC), mucinous ovarian carcinoma (MUOC), and low-grade serous ovarian carcinoma (LGSOC).

Duplicate cases (two samples with the same ottaID) were removed before generating the final training set to use for fitting the classification models. All CS3 cases were preferred over CS1

and CS2, and CS3-Vancouver cases were preferred over CS3-AOC and CS3-USC when selecting duplicates.

The final training set used only genes that were common across all three CodeSets.

Figure 2.2: Cohorts Selection

2.2 Classifiers

We used a supervised learning framework to train four multinomial classification algorithms in the Training Set. The full prediction model pipeline was run using SLURM batch jobs submitted to a partition on a Rocky 9 server. All resampling techniques, pre-processing, model specification, hyperparameter tuning, and evaluation metrics were implemented using the tidymodels suite of packages in R (v4.5.2). The classifiers we used are:

- Random Forest (`rf`)
- Support Vector Machine (`svm`)
- XGBoost (`xgb`)
- Regularized Multinomial Regression (`mr`)

2.2.1 Resampling of Training Set

We used a nested cross-validation design to evaluate each classifier and perform hyperparameter tuning. An outer 5-fold CV stratified by histotype, along with an inner 5-fold CV with two repeats also stratified by histotype, ensured that the test sets of the inner dataset contained a reasonable number of cases from the smallest minority class.

The hyperparameter combination that yielded the highest average F1 score across the inner training sets was selected for assessing prediction performance in the outer training loop. We avoided using the bootstrap method for outer resampling to prevent overlap between the training and test sets caused by sampling with replacement. This overlap could lead to inflated performance estimates, as some observations would be used for both training and evaluation in the inner loop.

2.2.2 Hyperparameter Tuning

The following specifications for each classifier were used for tuning hyperparameters:

- **rf:** The number of trees was fixed at 500. Two other hyperparameters were tuned across 10 randomly selected points in a space-filling parameter grid design:
  - mtry: number of randomly selected predictors
  - min_n: minimal node size
- **xgb**: The number of trees was fixed at 500. Seven other hyperparameters were tuned across 10 randomly selected points in a space-filling parameter grid design:
  - mtry: number of randomly selected predictors
  - min_n: minimal node size
  - tree_depth: maximum depth of the tree (splits)
  - learn_rate: rate at which the boosting algorithm adapts from iteration-to-iteration (shrinkage parameter)
  - loss_reduction: reduction in the loss function required to split further
  - sample_size: proportion of observations exposed to fitting routine
  - stop_iter: number of iterations without improvement before stopping
- **svm**: The cost and sigma hyperparameters were tuned across 10 randomly selected points in a space-filling parameter grid design. We tuned the cost parameter in the range [1, 8]. The range for tuning the sigma parameter was obtained from the 10% and 90% quantiles of the estimates from the kernlab::sigest() function.
- **mr**: The penalty (lambda) and mixture (alpha) hyperparameters were tuned across 100 randomly selected points in a space-filling parameter grid design.

The hyperparameter combination that resulted in the highest average F1-score across the inner training sets was selected for each classifier to use as the model for assessing prediction performance in the outer training loop.

2.2.3 Subsampling

One of the main concerns in the analysis is that the distribution of the five main histotypes is highly imbalanced, which has implications for the accuracy of the classifier. To mitigate this, we investigated several strategies based on subsampling approaches:

- **None**: No subsampling is performed
- **Down-sampling**: All levels except the minority class are sampled down to the same frequency as the minority class
- **Up-sampling**: All levels except the majority class are sampled up to the same frequency as the majority class
- **SMOTE**: All levels except the majority class have synthetic data generated until they have the same frequency as the majority class
- **Hybrid**: All levels except the majority class have synthetic data generated up to 50% of the frequency of the majority class, then the majority class is sampled down to the same frequency as the rest.

Figure 2.3 helps visualize how the distribution of classes changes when we apply subsampling techniques to handle class imbalance:

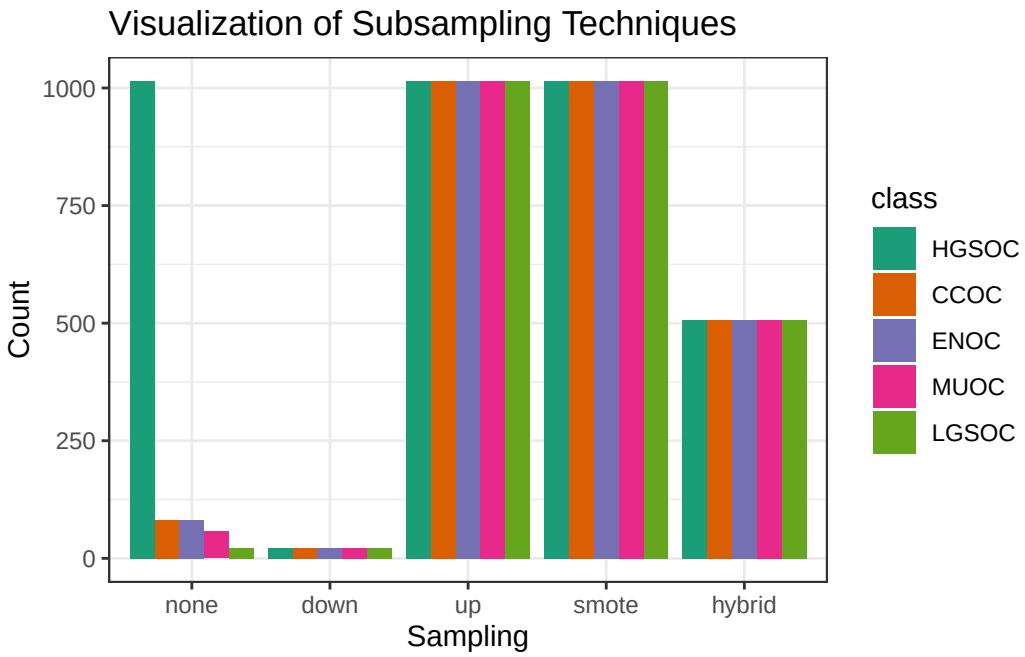


Figure 2.3: Visualization of Subsampling Techniques

All subsampling methods to mitigate class imbalance were applied within the inner CV loop.

2.2.4 Workflows

The four **algorithms** and five **subsampling** methods created 20 unique classification **workflows**. For example, the `hybrid_xgb` workflow pre-processes the training set by applying a hybrid sub-

sampling method and subsequently uses the XGBoost algorithm to classify ovarian histotypes. We additionally tested two ensemble algorithms, the two-step, and the sequential algorithms described in the following sections.

2.3 Two-Step Algorithm

The HGSOC histotype comprises most cases among ovarian carcinoma patients, while the remaining patients are uniformly split across ENOC, CCOC, MUOC, and LGSOC histotypes. We can implement a two-step algorithm as such:

- Step 1: Select the best classifier to predict HGSOC vs. non-HGSOC, using the workflow with the best per-class F1-score for HGSOC when fit to a 5-class prediction model
- Step 2: remove HGSOCs from the dataset and re-train a 4-class multinomial classifier to predict the four remaining non-HGSOC histotypes. Select the workflow with the best macro-averaged F1-score when fit to a 4-class prediction model.

Let

$$\begin{aligned} X_k &= \text{Training data with k classes} \\ C_k &= \text{Class with highest } F_1 \text{ score from training } X_k \\ W_k &= \text{Workflow associated with } C_k \end{aligned} \tag{2.5}$$

Figure 2.4 shows how the two-step algorithm works:

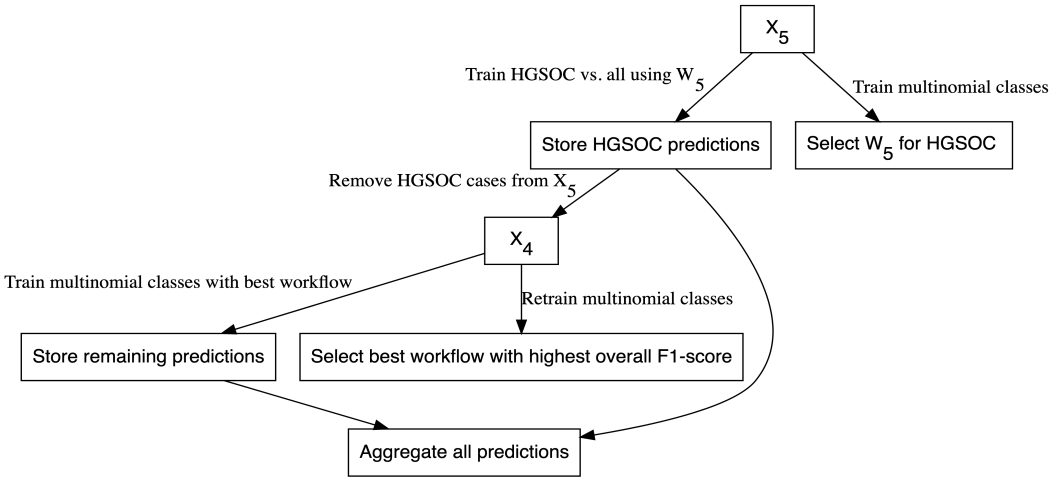


Figure 2.4: Two-Step Algorithm

2.3.1 Aggregating Predictions

The aggregation for two-step predictions is quite straightforward:

1. Predict HGSOC vs. non-HGSOC

2. Among all non-HGSOC cases, predict CCOC vs. LGSOC vs. MUOC vs. ENOC

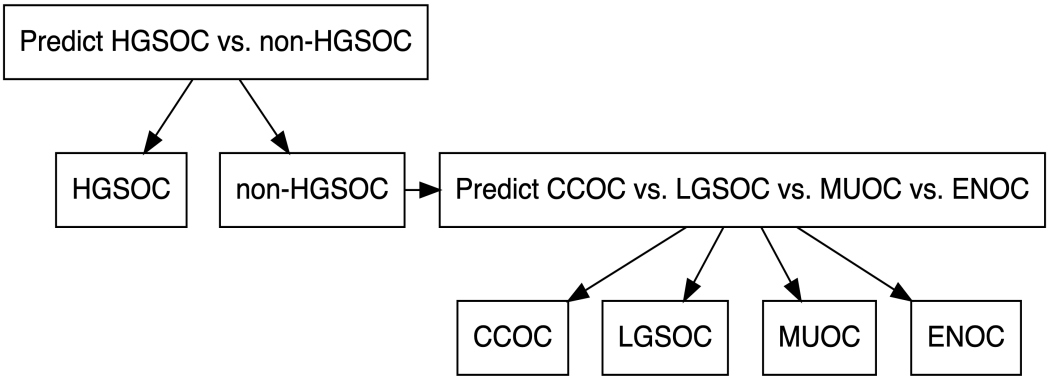


Figure 2.5: Aggregating Predictions for Two-Step Algorithm

2.4 Sequential Algorithm

Instead of training on k classes simultaneously using multinomial classifiers, we can use a sequential algorithm that iteratively performs one-vs-all binary classifications k-1 times to obtain a final prediction of all cases. At each step in the sequence, we classify one class vs. all other classes, where the classes that make up the "other" class are those not equal to the current "one" class and exclude all "one" classes from previous steps. For example, if the "one" class in step 1 was HGSOC, the "other" classes would include CCOC, ENOC, LGSOC, and MUOC. If the "one" class in step 2 was CCOC, the "other" classes would include ENOC, LGSOC, and MUOC.

The order of classes and workflows to use at each step in the sequential algorithm is determined after removing the data associated with an already classified class and retraining using the remaining data multinomial classifiers, as described before. The class and workflow to use for the next step in the sequence are selected based on the best per-class evaluation metric value (e.g., F1-score).

Figure 2.6 illustrates how the sequential algorithm works for K=5, using ovarian histotypes as an example for the classes.

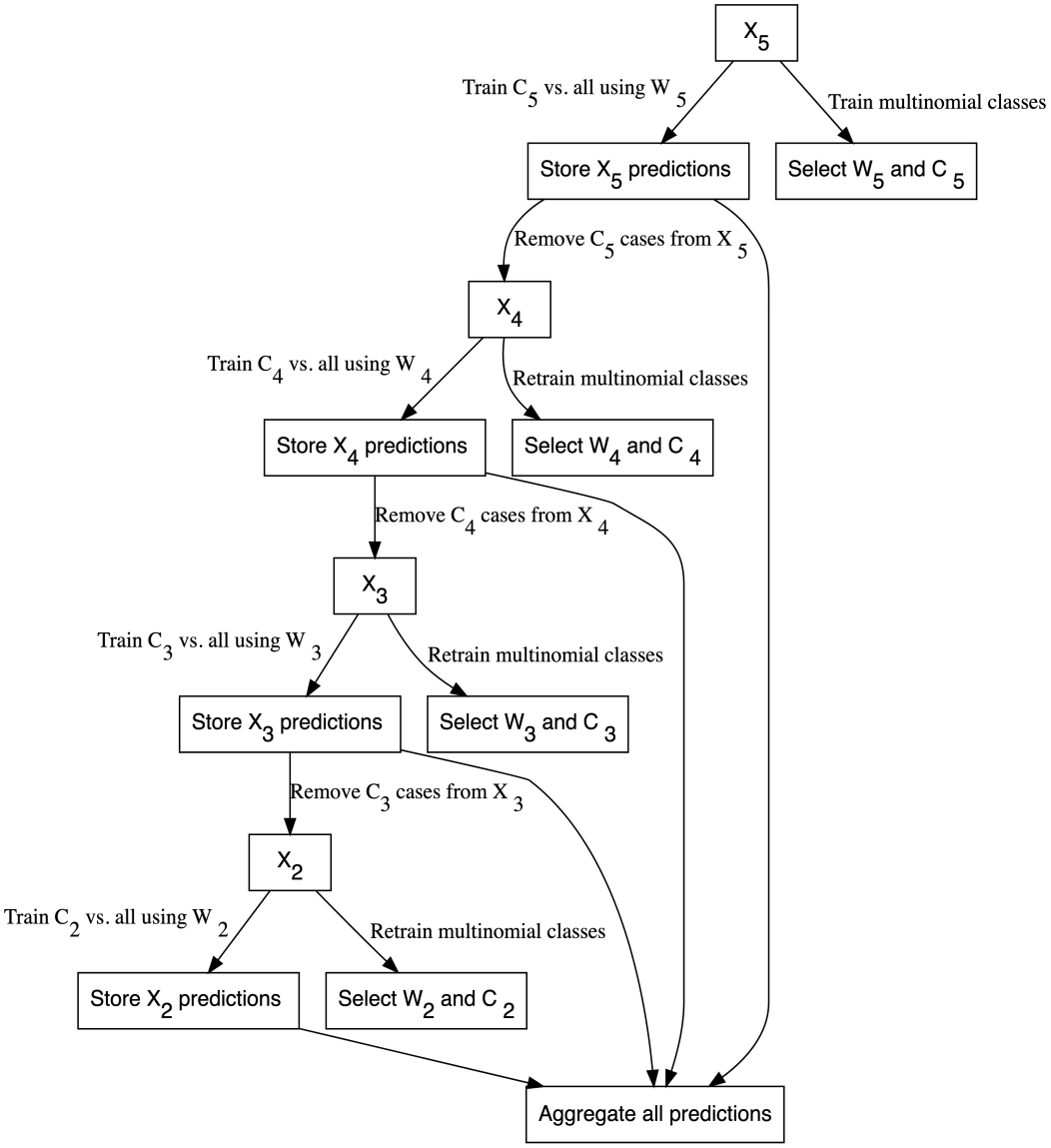


Figure 2.6: Sequential Algorithm

2.4.1 Aggregating Predictions

We have to aggregate the one-vs-all predictions from each of the sequential algorithm workflows in order to obtain a final class prediction on a holdout test set. Each sequential workflow has to be assessed on every sample to ensure that cases classified into the "other" class from a previous step of the sequence are eventually assigned to a predicted class. For example, if based on certain class-specific metrics, we determined that the order of classes in the sequential algorithm was to predict HGSC vs. non-HGSC, CCOC vs. non-CCOC, LGSOC vs. non-LGSOC, and then MUOC vs. ENOC. Figure 2.7 illustrates how the final predictions are assigned:

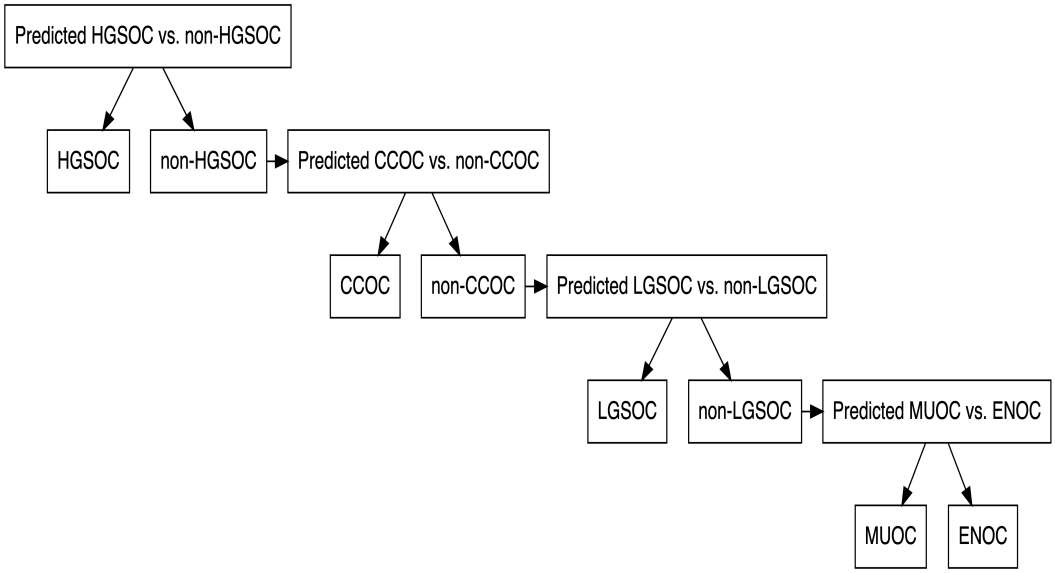


Figure 2.7: Aggregating Predictions for Sequential Algorithm

2.5 Performance Evaluation

2.5.1 Class Metrics

We use accuracy, sensitivity, specificity, F1-score, kappa, and balanced accuracy, as class metrics to measure both training and test performance between different workflows. Multiclass extensions of these metrics can be calculated except for F1-score, where we use macro-averaging to obtain an overall metric. Class-specific metrics are calculated by recoding classes into one-vs-all categories for each class.

2.5.1.1 Accuracy

The accuracy is defined as the proportion of correct predictions out of all cases:

$$\text{accuracy} = \frac{TP}{TP + FP + FN + TN} \tag{2.6}$$

2.5.1.2 Sensitivity

Sensitivity is the proportional of correctly predicted positive cases, out of all cases that were truly positive

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{2.7}$$

2.5.1.3 Specificity

Specificity is the proportional of correctly predicted negative cases, out of all cases that were truly negative.

$$\text{specificity} = \frac{TN}{TN + FP} \tag{2.8}$$

2.5.1.4 F1-Score

The F-measure can be thought of as a harmonic mean between precision and recall:

$$F_{meas} = \frac{(1 + \beta^2) \times precision \times recall}{(\beta^2 \times precision) + recall} \tag{2.9}$$

The β value can be adjusted to place more weight upon precision or recall. The most common value is β is 1, which is also commonly known as the F1-score. A multiclass extension doesn't exist for the F1-score, so we use macro-averaging to calculate this metric when there are more than two classes. For example, with k classes, the macro-averaged F1-score is equal to:

$$F_{1_{macro}} = \frac{1}{k} \sum_{i=1}^{k} F_{1_i} \tag{2.10}$$

where each F_{1_i} is the F1-score computed frrom recoding classes into $k = i$ vs. $k \neq i$.

In situations where there is not at least one predicted case for each of the classes (e.g. for a poor classifier), F_{1_i} is undefined because the per-class precision of class i is undefined. Those F_{1_i} terms are removed from the $F_{1_{macro}}$ equation and the resulting value may be inflated. Interpreting the F1-score in such a case would be misleading.

2.5.1.5 Balanced Accuracy

Balanced accuracy is the arithmetic mean of sensitivity and specificity.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{2.11}$$

2.5.1.6 Kappa

Kappa is the defined as:

$$\text{kappa} = \frac{p_0 - p_e}{1 - p_e} \tag{2.12}$$

where p_0 is the observed agreement among raters and p_e is the hypothetical probability of agreement due to random chance.

2.5.2 AUC

The area under the receiver operating curve (AUC) is calculated by adding up the area under the curve formed by plotting sensitivity vs. 1 - specificity. The Hand-till method is used as a multiclass extension for the AUC.

We did not use AUC to measure class-specific training set performance because combining predicted probabilities in a one-vs-all fashion might be potentially misleading. The sum of probabilities that add up to the "other" class is not equivalent to the predicted probability of the "other" class when using a multiclass classifier.

Instead, we only reported ROC curves and their associated AUCs for test set performance among the highest ranked algorithms.

2.6 Rank Aggregation

To select the best algorithm, we implemented a two-stage rank aggregation procedure using the Genetic Algorithm (Pihur et al. 2009). First, we ranked all workflows based on per-class F1-scores, balanced accuracy, and kappa to see which workflows performed well in predicting all five histotypes. Then, we took the ranks from these three performance metrics and performed a second run of rank aggregation. The top 5 workflows were determined from the final rank aggregation result.

2.7 Gene Optimization

We want to discover an optimal set of genes for the classifiers while including specific genes from other studies such as PrOTYPE and SPOT. A total of 72 genes are used in the classifier training set.

The classifier set includes 16 genes that overlap with the PrOTYPE classifier: COL11A1, CD74, CD2, TIMP3, LUM, CYTIP, COL3A1, THBS2, TCF7L1, HMGA2, FN1, POSTN, COL1A2, COL5A2, PDZK1IP1, FBN1.

The classifier set also includes 13 genes that overlap with the SPOT signature: HIF1A, CXCL10, DUSP4, SOX17, MITF, CDKN3, BRCA2, CEACAM5, ANXA4, SERPINE1, TCF7L1, CRABP2, DNAJC9.

We obtain 28 genes from the union of PrOTYPE and SPOT genes that we want to include in the final classifier, regardless of model performance. We then incrementally add genes one at a time from the remaining 44 candidate genes based on a variable importance rank to the set of 28 base genes and recalculate performance metrics. The optimal number of genes is determined by the number of genes needed to achieve the highest average F1-Score across classes, including the overall metric in the average. We use an F1-score that is averaged across cross-validation folds (5) and class groups (6: Overall, HGSOC, CCOC, ENOC, MUOC, LGSOC) to compare performance between different number of genes selected.

Here is the breakdown of genes used and whether they belong to the PrOTYPE and/or SPOT sets:

Table 2.1: Gene Distribution

Genes	PrOTYPE	SPOT
TCF7L1	v	v
COL11A1	v	
CD74	v	
CD2	v	
TIMP3	v	
LUM	v	
CYTIP	v	
COL3A1	v	
THBS2	v	
HMGA2	v	
FN1	v	
POSTN	v	
COL1A2	v	
COL5A2	v	
PDZK1IP1	v	
FBN1	v	
HIF1A		v
CXCL10		v
DUSP4		v
SOX17		v
MITF		v
CDKN3		v
BRCA2		v
CEACAM5		v
ANXA4		v
SERPINE1		v
CRABP2		v
DNAJC9		v
C10orf116		
GAD1		
TPX2		
KGFLP2		
EGFL6		
KLK7		
PBX1		
LIN28B		
TFF3		
MUC5B		
FUT3		
STC1		
BCL2		
PAX8		
GCNT3		
GPR64		

ADCYAP1R1
IGKC
BRCA1
IGJ
TFF1
MET
CYP2C18
CYP4B1
SLC3A1
EPAS1
HNF1B
IL6
ATP5G3
DKK4
SENP8
CAPN2
C1orf173
CPNE8
IGFBP1
WT1
TP53
SEMA6A
SERPINA5
ZBED1
TSPAN8
SCGB1D2
LGALS4
MAP1LC3A

2.7.1 Variable Importance

Variable importance is calculated using either a model-based approach if it is available, or a permutation-based VI score otherwise (e.g., for SVM). The variable importance scores are averaged across the outer training folds, and then ranked from highest to lowest.

For the sequential and two-step classifiers, we calculate an overall VI rank by taking the cumulative union of genes at each variable importance rank across all sequences, until all genes have been included.

The variable importance measures are:

- Random Forest: impurity measure (Gini index)
- XGBoost: gain (fractional contribution of each feature to the model based on the total gain of the corresponding features's splits)
- SVM: permutation based p-values

- Multinomial regression: absolute value of estimated coefficients at cross-validated lambda value

3 Distributions

3.1 Histotype Distribution

Table 3.1: Histotype Distribution in Training Set by Processing Stage

Variable	Levels	CS3	CS1	CS2	Total
Selected Cohorts					
Histotype	HGSOC	1808 (78%)	128 (47%)	655 (78%)	2591 (75%)
	CCOC	164 (7%)	48 (18%)	62 (7%)	274 (8%)
	ENOC	250 (11%)	60 (22%)	49 (6%)	359 (10%)
	MUOC	68 (3%)	17 (6%)	58 (7%)	143 (4%)
	LGSOC	36 (2%)	19 (7%)	20 (2%)	75 (2%)
	Missing	151	22	59	232
Total	N (%)	2477 (67%)	294 (8%)	903 (25%)	3674 (100%)
QC					
Histotype	HGSOC	1676 (78%)	122 (46%)	641 (78%)	2439 (75%)
	CCOC	158 (7%)	48 (18%)	62 (8%)	268 (8%)
	ENOC	213 (10%)	60 (23%)	47 (6%)	320 (10%)
	MUOC	65 (3%)	16 (6%)	56 (7%)	137 (4%)
	LGSOC	36 (2%)	18 (7%)	20 (2%)	74 (2%)
	Missing	125	22	56	203
Total	N (%)	2273 (66%)	286 (8%)	882 (26%)	3441 (100%)
Main Histotypes					
Histotype	HGSOC	1676 (78%)	122 (46%)	641 (78%)	2439 (75%)
	CCOC	158 (7%)	48 (18%)	62 (8%)	268 (8%)
	ENOC	213 (10%)	60 (23%)	47 (6%)	320 (10%)
	MUOC	65 (3%)	16 (6%)	56 (7%)	137 (4%)
	LGSOC	36 (2%)	18 (7%)	20 (2%)	74 (2%)
Total	N (%)	2148 (66%)	264 (8%)	826 (26%)	3238 (100%)
Removed Duplicates					
	HGSOC	1578 (78%)	118 (48%)	623 (78%)	2319 (76%)

Histotype	CCOC	146 (7%)	45 (18%)	56 (7%)	247 (8%)
	ENOC	200 (10%)	56 (23%)	43 (5%)	299 (10%)
	MUOC	55 (3%)	13 (5%)	54 (7%)	122 (4%)
	LGSOC	32 (2%)	14 (6%)	19 (2%)	65 (2%)
Total	N (%)	2011 (66%)	246 (8%)	795 (26%)	3052 (100%)
Normalized and Recombined					
Histotype	HGSOC	454 (97%)	117 (49%)	622 (79%)	1193 (79%)
	CCOC	4 (1%)	44 (18%)	55 (7%)	103 (7%)
	ENOC	4 (1%)	55 (23%)	42 (5%)	101 (7%)
	MUOC	4 (1%)	12 (5%)	53 (7%)	69 (5%)
	LGSOC	4 (1%)	13 (5%)	18 (2%)	35 (2%)
Total	N (%)	470 (31%)	241 (16%)	790 (53%)	1501 (100%)
Removed Replicates					
Histotype	CCOC	4 (50%)	24 (39%)	53 (57%)	81 (50%)
	ENOC	4 (50%)	38 (61%)	40 (43%)	82 (50%)
	Missing	462	15	617	1094
Total	N (%)	470 (37%)	77 (6%)	710 (56%)	1257 (100%)

Table 3.2: Histotype Distribution in Training, Confirmation, and Validation Sets

Variable	Levels	Training	Confirmation	Validation
Histotype	CCOC	81 (50%)	72 (40%)	69 (44%)
	ENOC	82 (50%)	107 (60%)	88 (56%)
	Missing	1094	463	737
Total	N (%)	1257 (45%)	642 (23%)	894 (32%)

3.2 Cohort Distribution

Table 3.3: Pre-QC Cohort Distribution by CodeSet

CodeSet	CS1 N = 294	CS2 N = 903	CS3 N = 2,477
Cohort			
OOU	108 (37%)	43 (4.8%)	19 (0.8%)
OOUE	32 (11%)	30 (3.3%)	11 (0.4%)
VOA	145 (49%)	122 (14%)	538 (22%)
OVAR3	0 (0%)	150 (17%)	0 (0%)
OVAR11	0 (0%)	416 (46%)	0 (0%)
MAYO	6 (2.0%)	63 (7.0%)	0 (0%)
DOVE4	0 (0%)	0 (0%)	1,160 (47%)
TNCO	0 (0%)	0 (0%)	691 (28%)
MTL	3 (1.0%)	59 (6.5%)	0 (0%)
JAPAN	0 (0%)	8 (0.9%)	0 (0%)
POOL-CTRL	0 (0%)	12 (1.3%)	0 (0%)
POOL-1	0 (0%)	0 (0%)	31 (1.3%)
POOL-2	0 (0%)	0 (0%)	14 (0.6%)
POOL-3	0 (0%)	0 (0%)	13 (0.5%)

[1] n (%)

3.3 Quality Control

3.3.1 Failed Samples

We use an aggregated `QCFlag` that considers a sample to have failed QC if any of the following QC conditions are flagged:

- Linearity
- Imaging
- Smallest Positive Control
- Normality

Table 3.4: Quality Control Summary

Quality Control Flag	Training N = 1,753	Confirmation N = 691	Validation N = 1,160	Total N = 3,604
Linearity	4 (0.2%)	0 (0%)	0 (0%)	4 (0.1%)
Imaging	3 (0.2%)	0 (0%)	4 (0.3%)	7 (0.2%)
Smallest Positive Control	2 (0.1%)	0 (0%)	0 (0%)	2 (<0.1%)
Normality	26 (1.5%)	1 (0.1%)	197 (17%)	224 (6.2%)
Overall QC	31 (1.8%)	1 (0.1%)	201 (17%)	233 (6.5%)

[1] n (%)

3.3.2 %GD vs. SNR

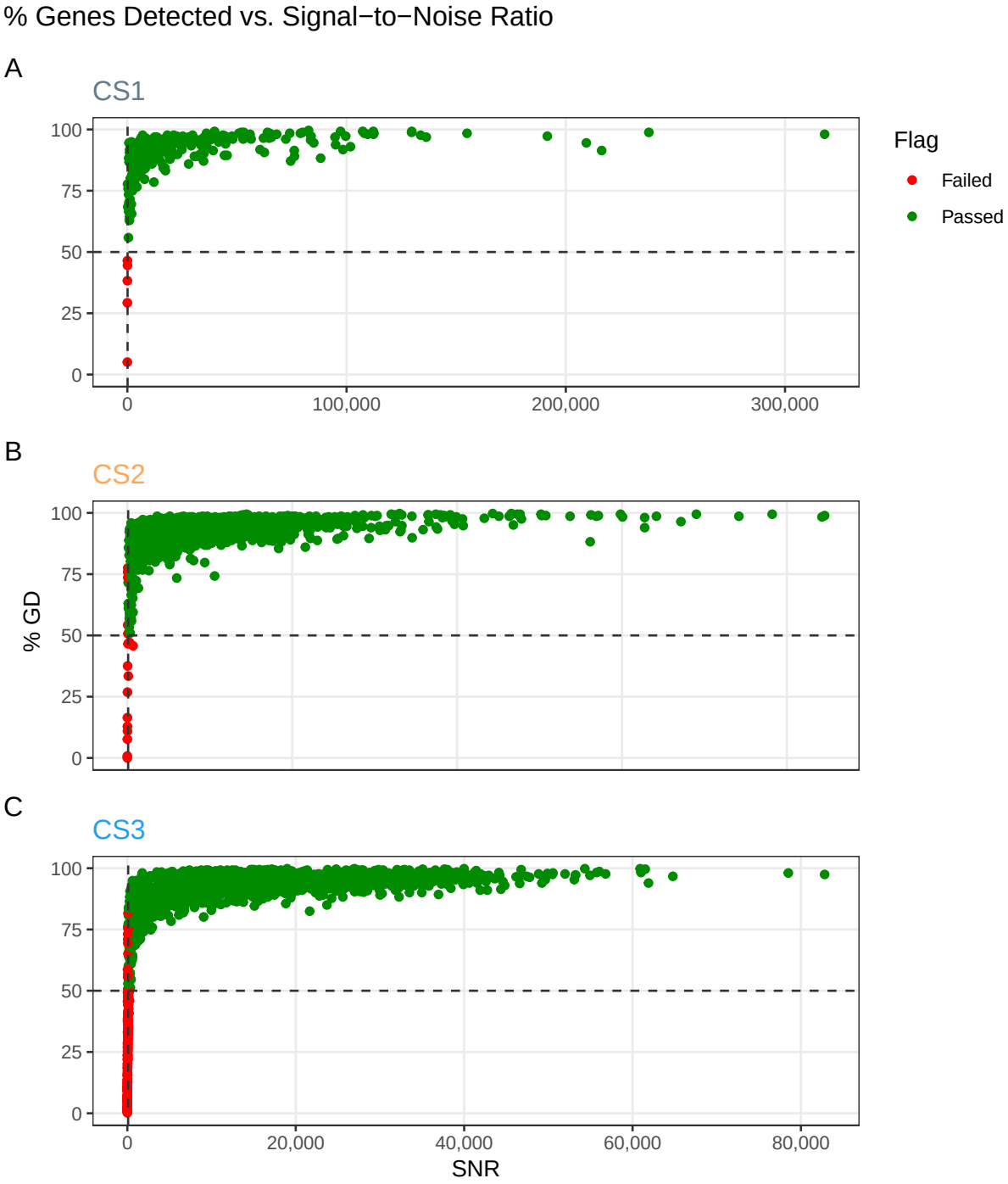


Figure 3.1: % Genes Detected vs. Signal to Noise Ratio

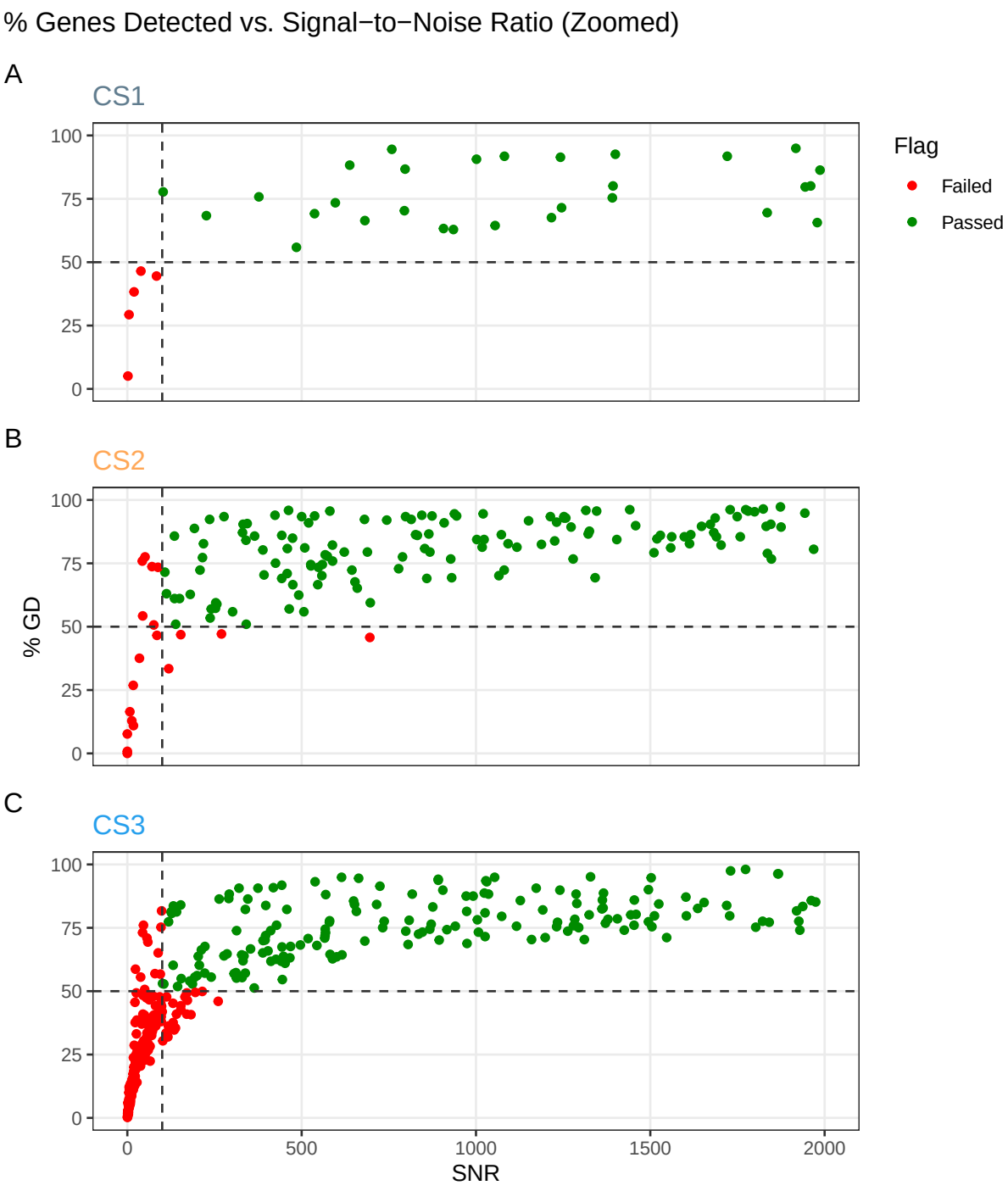


Figure 3.2: % Genes Detected vs. Signal to Noise Ratio (Zoomed)

3.4 Pairwise Gene Expression

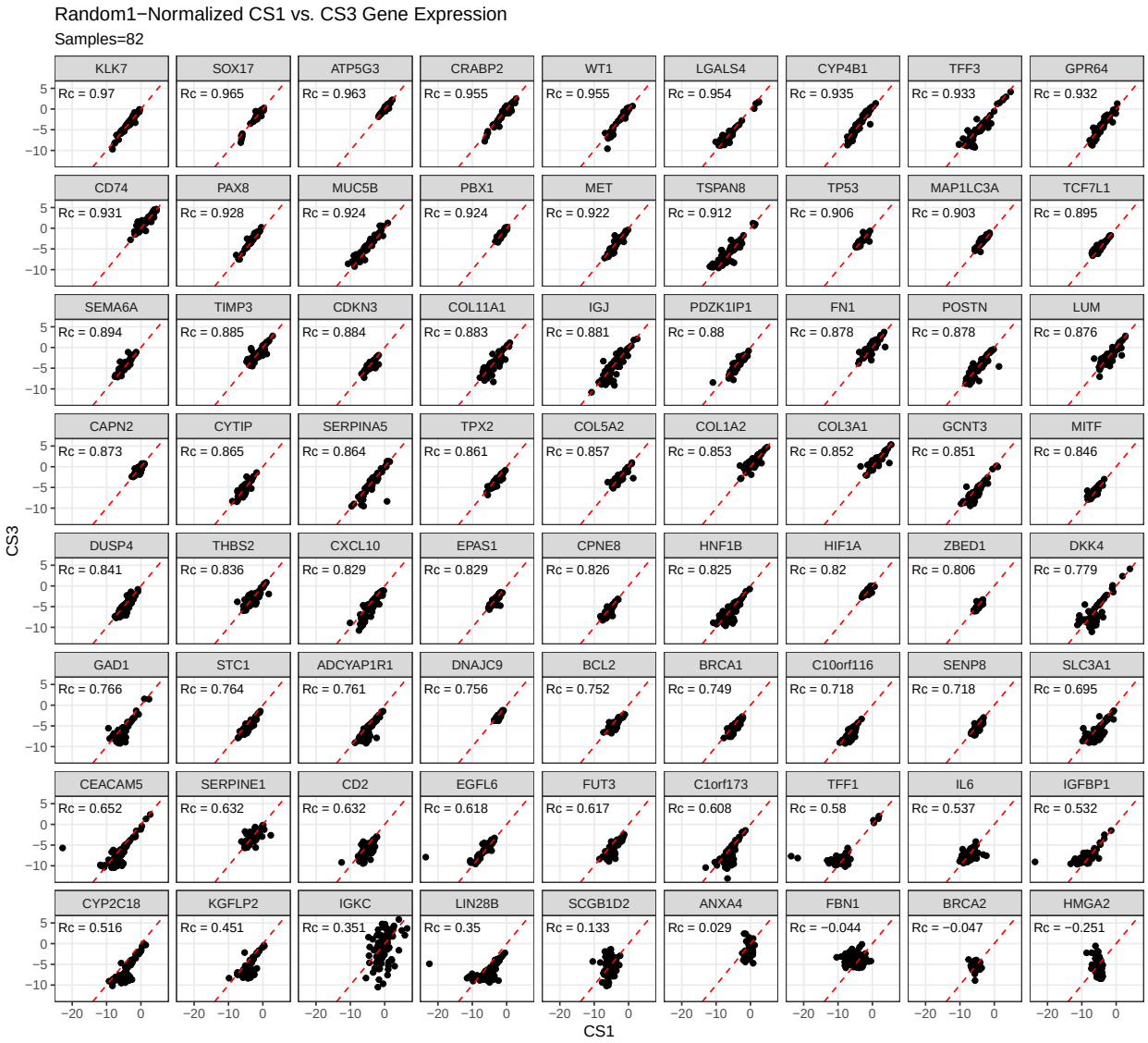


Figure 3.3: Random1-Normalized CS1 vs. CS3 Gene Expression

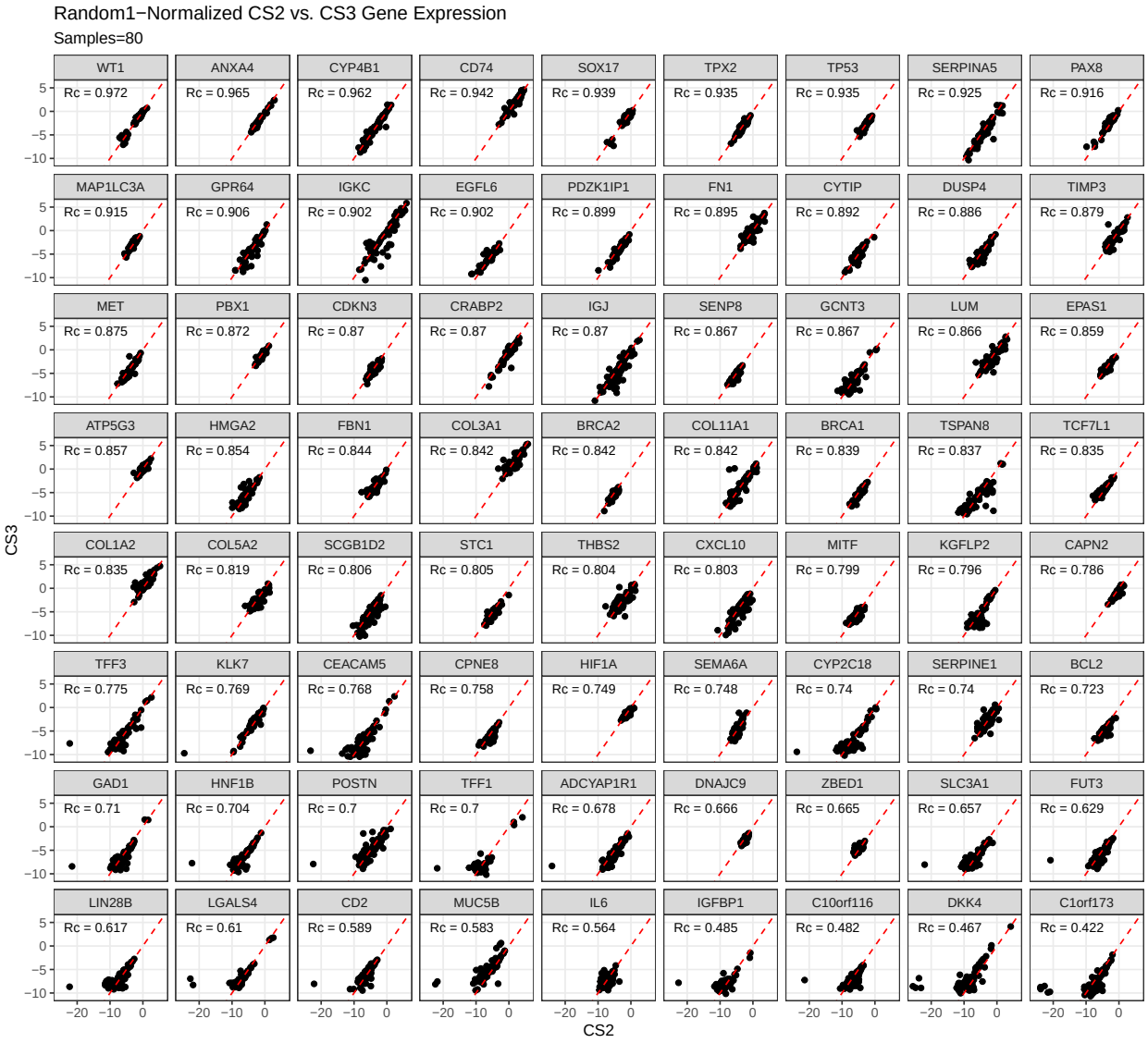


Figure 3.4: Random1-Normalized CS2 vs. CS3 Gene Expression

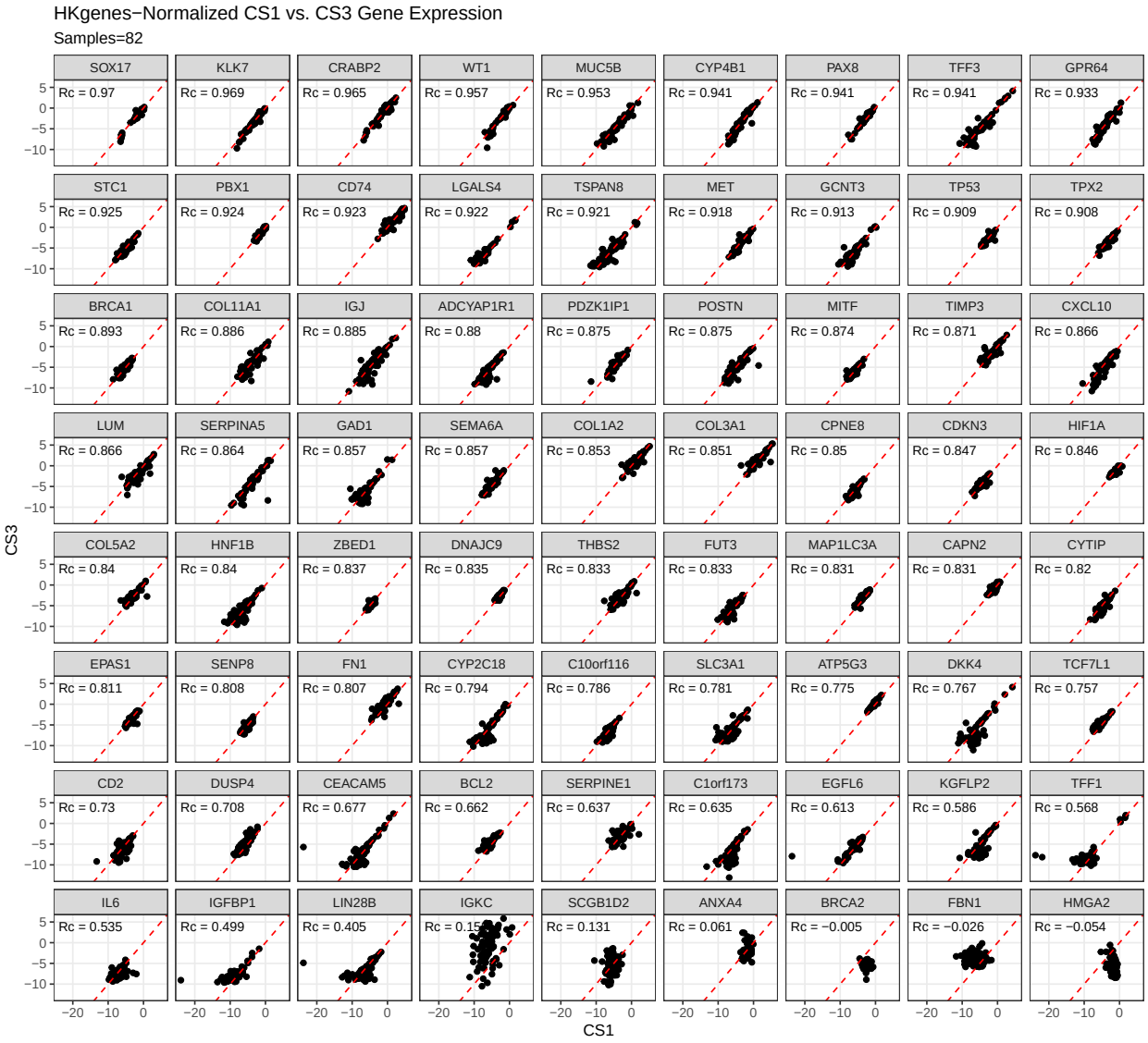


Figure 3.5: HKgenes-Normalized CS1 vs. CS3 Gene Expression

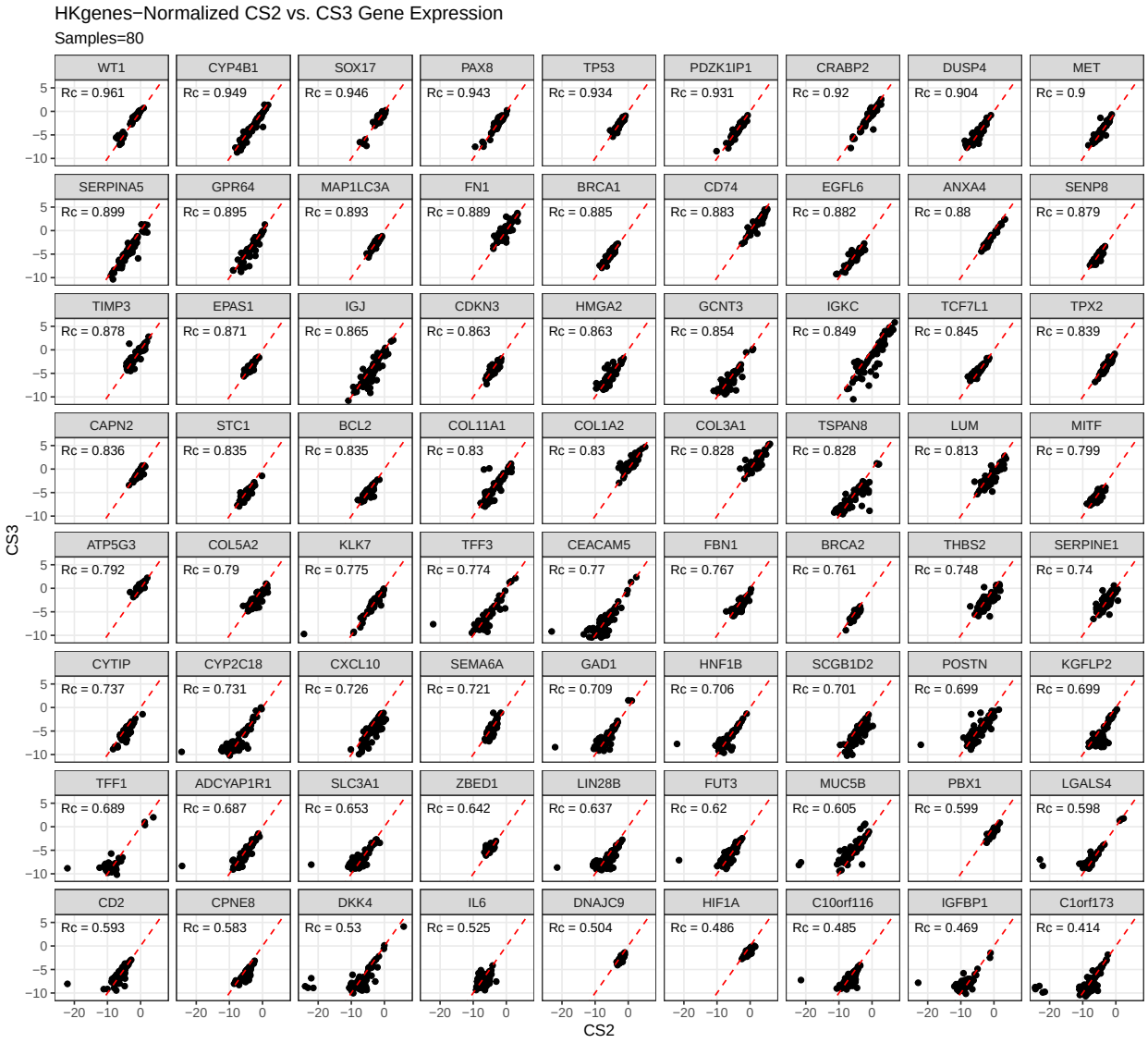


Figure 3.6: HKgenes-Normalized CS2 vs. CS3 Gene Expression

Table 3.5: Wilcoxon signed rank test of gene correlations between normalization methods

Correlation	**Housekeeping Genes** N = 72[1]	**Random1** N = 72[1]	**p-value**[2]
CS1 vs. CS3	0.84 (0.74, 0.90)	0.84 (0.67, 0.89)	0.079
CS2 vs. CS3	0.80 (0.69, 0.88)	0.84 (0.72, 0.88)	0.003

[1]Median (Q1, Q3)
[2]Wilcoxon signed rank test with continuity correction

4 Results

We summarize cross-validated training performance of class metrics in the training set. The accuracy, F1-score, and kappa, are the metrics of interest. Workflows are ordered by their mean estimates across the outer folds of the nested CV for each metric.

4.1 Training Set

4.1.1 Accuracy

Table 4.1: Training Set Mean Accuracy

			Histotypes				
Subsampling	Algorithms	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
none	rf	0.922	0.948	0.983	0.953	0.975	0.984
	svm	0.92	0.947	0.982	0.955	0.975	0.982
	xgb	0.807	0.809	0.936	0.933	0.955	0.982
	mr	0.807	0.807	0.936	0.935	0.955	0.982
down	rf	0.804	0.853	0.975	0.903	0.959	0.92
	svm	0.793	0.833	0.973	0.884	0.969	0.928
	xgb	0.783	0.833	0.968	0.897	0.961	0.907
	mr	0.829	0.87	0.977	0.912	0.961	0.938
up	rf	0.928	0.955	0.983	0.959	0.979	0.982
	svm	0.92	0.951	0.974	0.951	0.979	0.985
	xgb	0.928	0.959	0.982	0.96	0.974	0.982
	mr	0.891	0.924	0.976	0.949	0.966	0.967
smote	rf	0.921	0.955	0.98	0.953	0.97	0.984
	svm	0.92	0.95	0.979	0.954	0.978	0.981
	xgb	0.923	0.955	0.98	0.955	0.971	0.984
	mr	0.897	0.931	0.979	0.951	0.967	0.967
hybrid	rf	0.922	0.953	0.98	0.952	0.975	0.983
	svm	0.916	0.944	0.982	0.951	0.975	0.979
	xgb	0.924	0.954	0.984	0.958	0.971	0.982
	mr	0.892	0.925	0.978	0.947	0.967	0.966

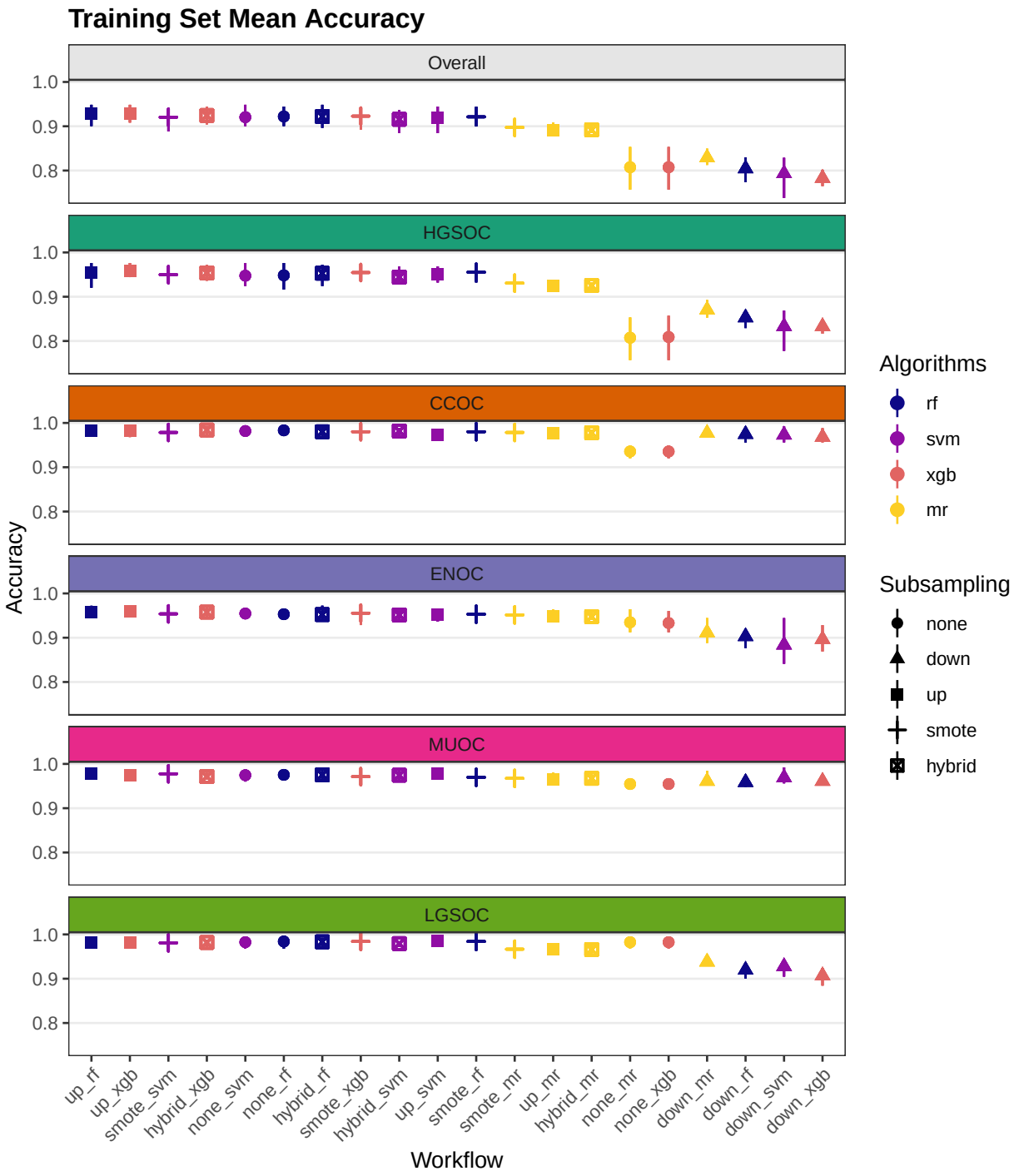


Figure 4.1: Training Set Mean Accuracy

4.1.2 Sensitivity

Table 4.2: Training Set Mean Sensitivity

			Histotypes				
Subsampling	Algorithms	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
none	rf	0.636	0.994	0.794	0.498	0.728	0.167
	svm	0.615	0.989	0.762	0.576	0.751	0
	xgb	0.2	1	0	0	0	0
	mr	0.2	1	0	0	0	0
down	rf	0.742	0.83	0.827	0.565	0.657	0.833
	svm	0.779	0.805	0.759	0.611	0.836	0.883
	xgb	0.714	0.802	0.843	0.589	0.67	0.667
	mr	0.776	0.848	0.831	0.62	0.77	0.808
up	rf	0.673	0.988	0.793	0.635	0.766	0.183
	svm	0.714	0.976	0.735	0.625	0.71	0.525
	xgb	0.718	0.981	0.805	0.624	0.781	0.4
	mr	0.788	0.921	0.849	0.666	0.786	0.717
smote	rf	0.658	0.984	0.748	0.618	0.75	0.192
	svm	0.759	0.966	0.757	0.681	0.748	0.642
	xgb	0.777	0.965	0.831	0.615	0.804	0.667
	mr	0.788	0.93	0.835	0.654	0.802	0.717
hybrid	rf	0.72	0.972	0.805	0.638	0.804	0.383
	svm	0.764	0.957	0.793	0.696	0.733	0.642
	xgb	0.77	0.966	0.851	0.624	0.818	0.592
	mr	0.783	0.923	0.835	0.676	0.767	0.717

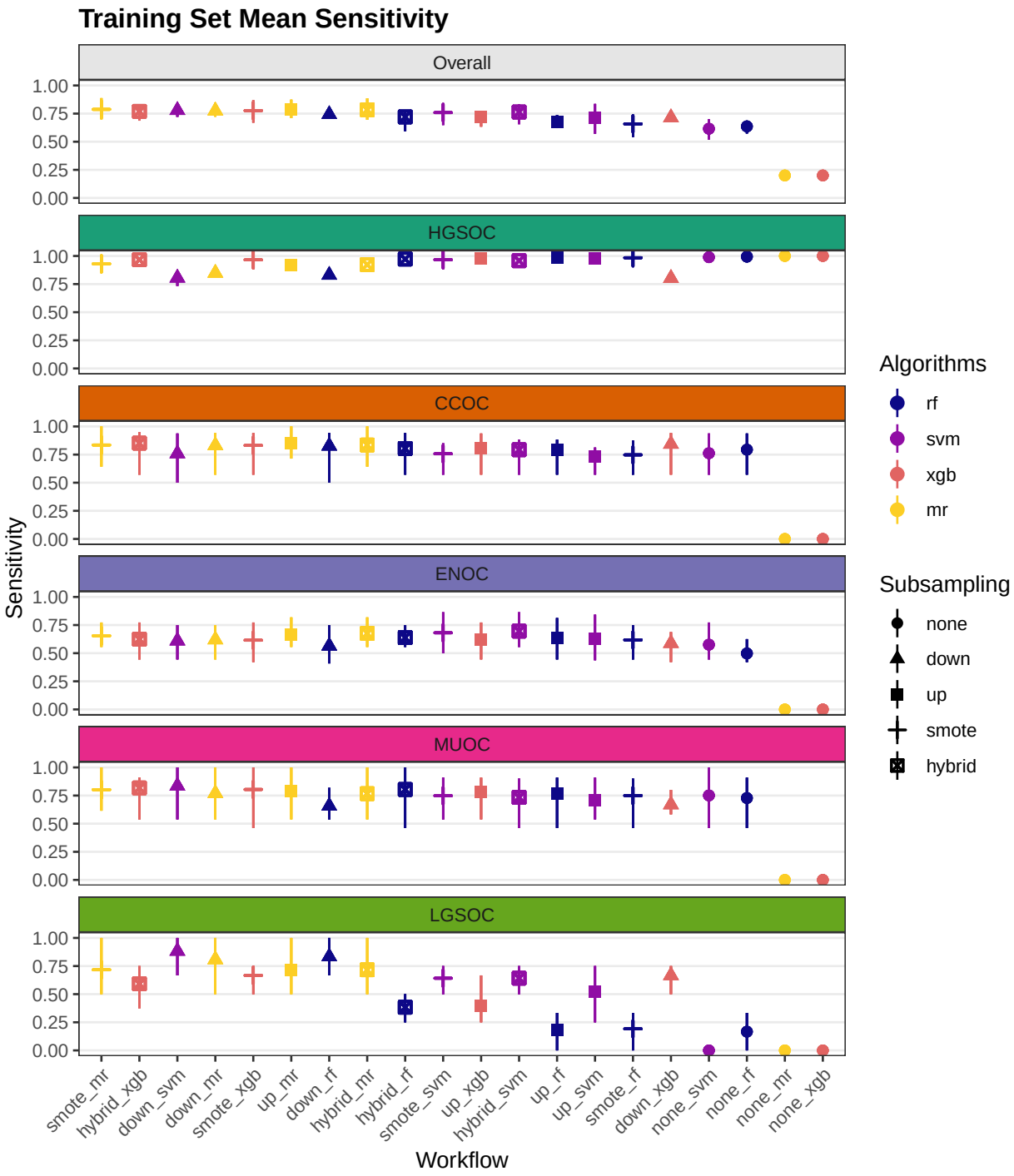


Figure 4.2: Training Set Mean Sensitivity

4.1.3 Specificity

Table 4.3: Training Set Mean Specificity

			Histotypes				
Subsampling	Algorithms	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
none	rf	0.946	0.764	0.996	0.985	0.988	0.999
	svm	0.947	0.775	0.997	0.98	0.986	1
	xgb	0.802	0.009	1	0.998	1	1
	mr	0.8	0	1	1	1	1
down	rf	0.952	0.955	0.984	0.928	0.973	0.921
	svm	0.949	0.952	0.987	0.902	0.976	0.928
	xgb	0.949	0.965	0.976	0.918	0.975	0.911
	mr	0.958	0.963	0.986	0.931	0.971	0.94
up	rf	0.957	0.822	0.996	0.98	0.989	0.997
	svm	0.958	0.844	0.99	0.973	0.992	0.993
	xgb	0.964	0.867	0.994	0.982	0.983	0.993
	mr	0.967	0.938	0.985	0.968	0.975	0.971
smote	rf	0.958	0.839	0.996	0.975	0.981	0.998
	svm	0.964	0.88	0.993	0.971	0.989	0.987
	xgb	0.969	0.91	0.99	0.978	0.98	0.99
	mr	0.968	0.934	0.988	0.971	0.976	0.971
hybrid	rf	0.964	0.874	0.992	0.974	0.984	0.994
	svm	0.965	0.89	0.994	0.968	0.987	0.985
	xgb	0.969	0.904	0.992	0.98	0.979	0.989
	mr	0.967	0.934	0.987	0.965	0.977	0.97

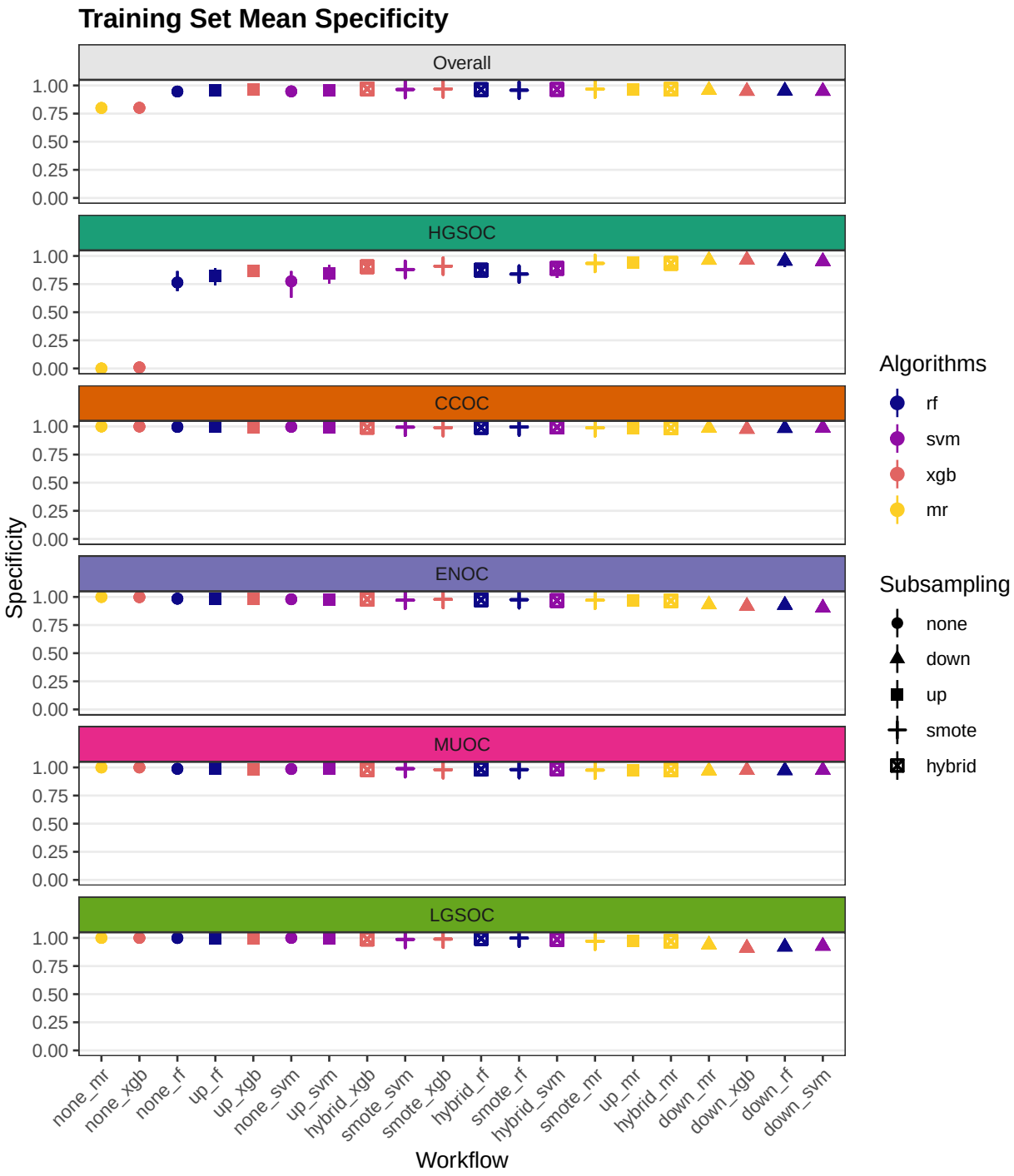


Figure 4.3: Training Set Mean Specificity

4.1.4 F1-Score

Table 4.4: Training Set Mean F1-Score

			Histotypes				
Subsampling	Algorithms	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
none	rf	0.733	0.968	0.848	0.568	0.723	0.4
	svm	0.782	0.968	0.834	0.607	0.719	NaN
	xgb	0.712	0.894	NaN	0	NaN	NaN
	mr	0.893	0.893	NaN	NaN	NaN	NaN
down	rf	0.594	0.901	0.799	0.419	0.589	0.262
	svm	0.617	0.886	0.778	0.412	0.713	0.295
	xgb	0.576	0.886	0.771	0.42	0.608	0.194
	mr	0.628	0.913	0.817	0.47	0.638	0.302
up	rf	0.755	0.972	0.849	0.652	0.757	0.362
	svm	0.725	0.97	0.778	0.605	0.75	0.521
	xgb	0.723	0.974	0.844	0.651	0.728	0.416
	mr	0.695	0.951	0.816	0.619	0.674	0.413
smote	rf	0.68	0.972	0.82	0.616	0.688	0.304
	svm	0.739	0.969	0.811	0.638	0.752	0.524
	xgb	0.744	0.972	0.833	0.628	0.71	0.579
	mr	0.704	0.956	0.827	0.629	0.691	0.418
hybrid	rf	0.724	0.971	0.829	0.631	0.741	0.45
	svm	0.733	0.965	0.838	0.635	0.72	0.507
	xgb	0.741	0.971	0.86	0.643	0.716	0.515
	mr	0.696	0.952	0.821	0.613	0.682	0.412

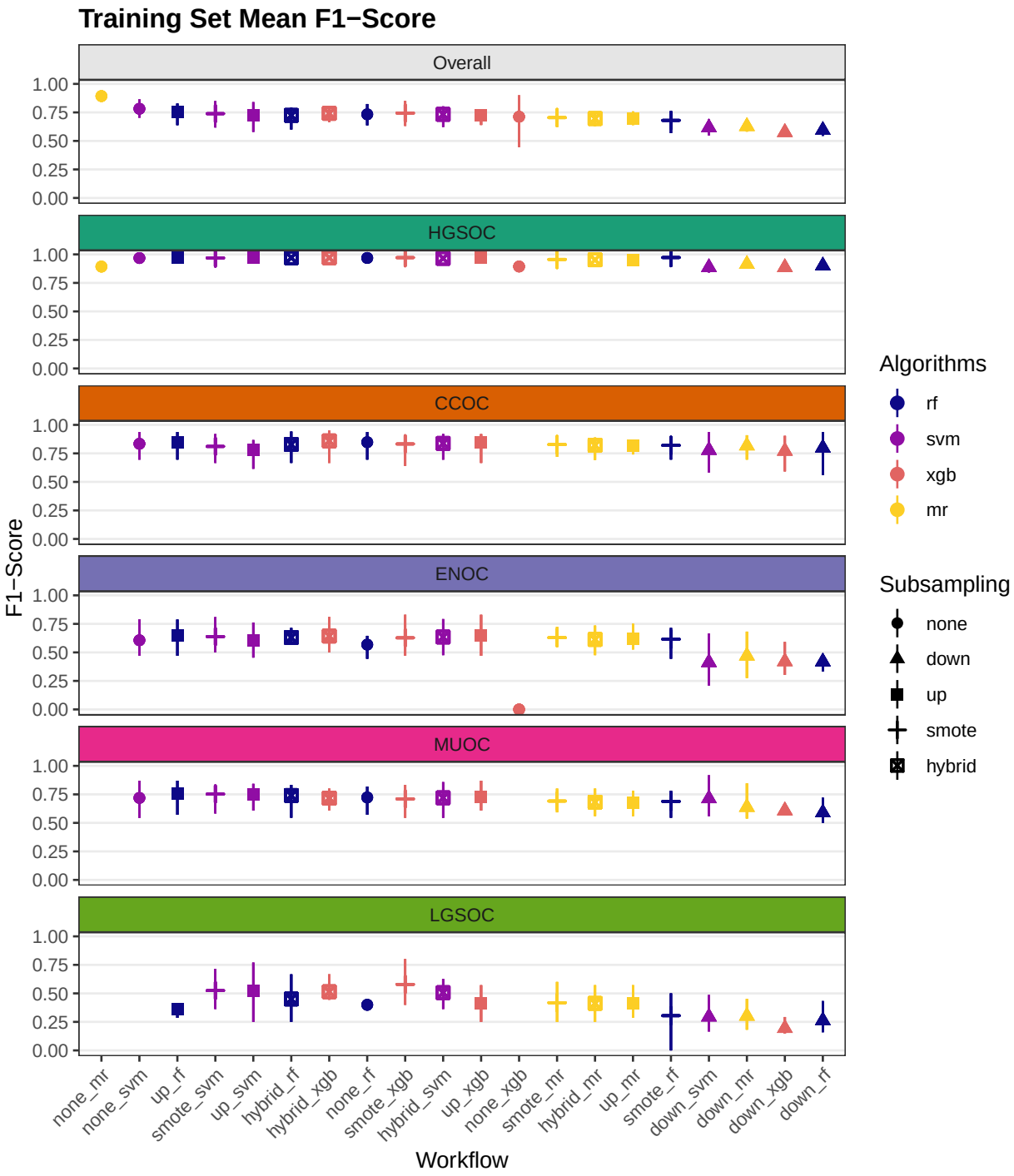


Figure 4.4: Training Set Mean F1-Score

4.1.5 Balanced Accuracy

Table 4.5: Training Set Mean Balanced Accuracy

			Histotypes				
Subsampling	Algorithms	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
none	rf	0.791	0.879	0.895	0.741	0.858	0.583
	svm	0.781	0.882	0.879	0.778	0.868	0.5
	xgb	0.501	0.505	0.5	0.499	0.5	0.5
	mr	0.5	0.5	0.5	0.5	0.5	0.5
down	rf	0.847	0.892	0.905	0.746	0.815	0.877
	svm	0.864	0.879	0.873	0.757	0.906	0.906
	xgb	0.832	0.883	0.91	0.753	0.822	0.789
	mr	0.867	0.906	0.909	0.776	0.87	0.874
up	rf	0.815	0.905	0.894	0.808	0.878	0.59
	svm	0.836	0.91	0.862	0.799	0.851	0.759
	xgb	0.841	0.924	0.9	0.803	0.882	0.696
	mr	0.878	0.93	0.917	0.817	0.881	0.844
smote	rf	0.808	0.911	0.872	0.797	0.865	0.595
	svm	0.862	0.923	0.875	0.826	0.869	0.814
	xgb	0.873	0.938	0.911	0.797	0.892	0.828
	mr	0.878	0.932	0.912	0.813	0.889	0.844
hybrid	rf	0.842	0.923	0.898	0.806	0.894	0.689
	svm	0.864	0.923	0.893	0.832	0.86	0.814
	xgb	0.87	0.935	0.922	0.802	0.899	0.791
	mr	0.875	0.928	0.911	0.821	0.872	0.843

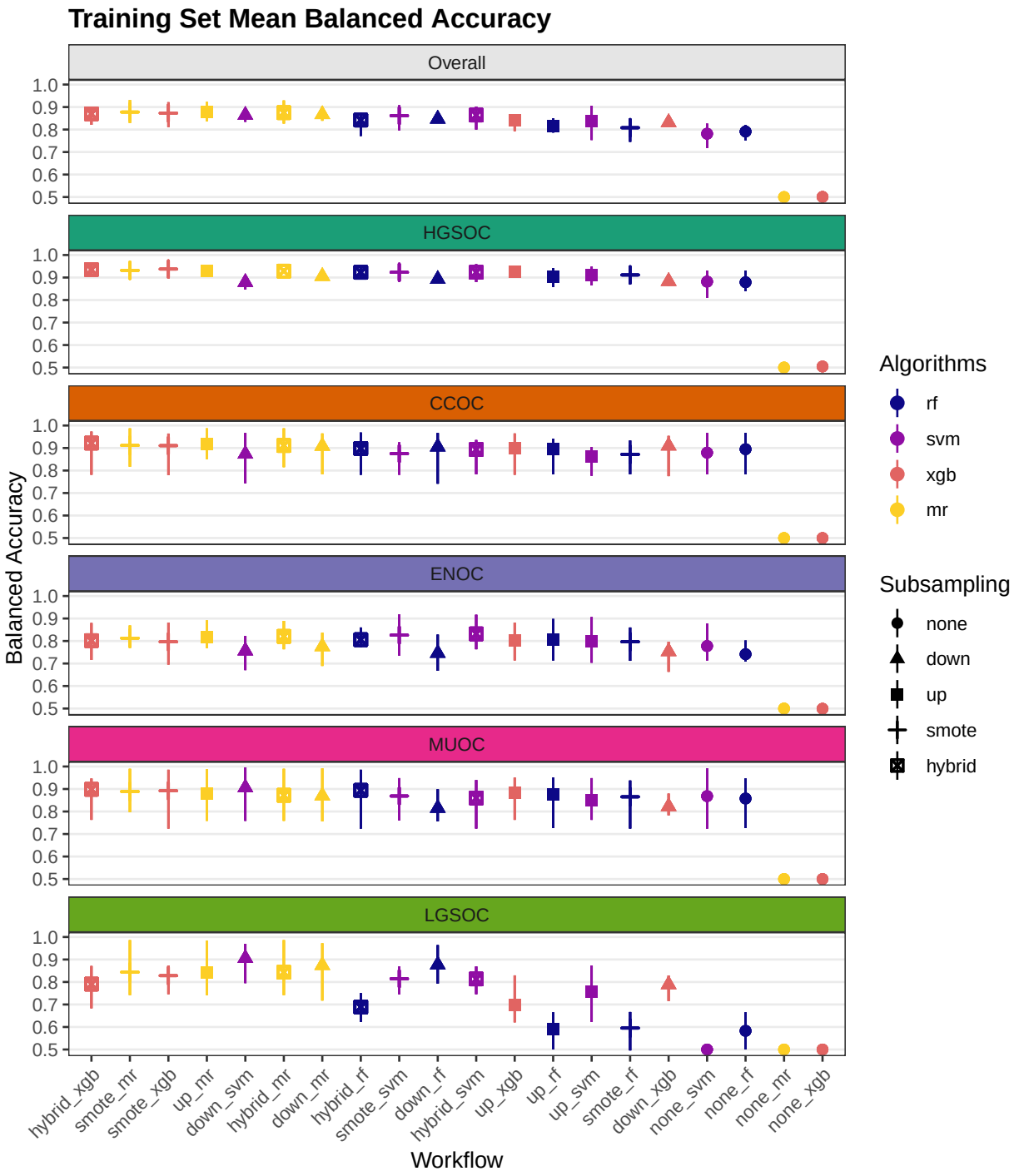


Figure 4.5: Training Set Mean Balanced Accuracy

4.1.6 Kappa

Table 4.6: Training Set Mean Kappa

			Histotypes				
Subsampling	Algorithms	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
none	rf	0.744	0.822	0.84	0.545	0.711	0.237
	svm	0.741	0.818	0.825	0.583	0.706	0
	xgb	0.007	0.016	0	-0.003	0	0
	mr	0	0	0	0	0	0
down	rf	0.55	0.621	0.785	0.37	0.568	0.24
	svm	0.539	0.584	0.764	0.359	0.697	0.275
	xgb	0.519	0.584	0.754	0.37	0.588	0.17
	mr	0.595	0.657	0.805	0.426	0.618	0.282
up	rf	0.776	0.85	0.84	0.63	0.746	0.213
	svm	0.749	0.835	0.764	0.579	0.739	0.513
	xgb	0.781	0.865	0.835	0.63	0.714	0.408
	mr	0.708	0.778	0.803	0.592	0.657	0.399
smote	rf	0.756	0.853	0.81	0.591	0.673	0.299
	svm	0.761	0.839	0.8	0.614	0.741	0.516
	xgb	0.773	0.856	0.822	0.604	0.695	0.571
	mr	0.72	0.793	0.816	0.603	0.675	0.405
hybrid	rf	0.767	0.85	0.818	0.605	0.728	0.443
	svm	0.753	0.825	0.828	0.609	0.707	0.497
	xgb	0.778	0.855	0.852	0.62	0.702	0.506
	mr	0.707	0.779	0.81	0.585	0.665	0.399

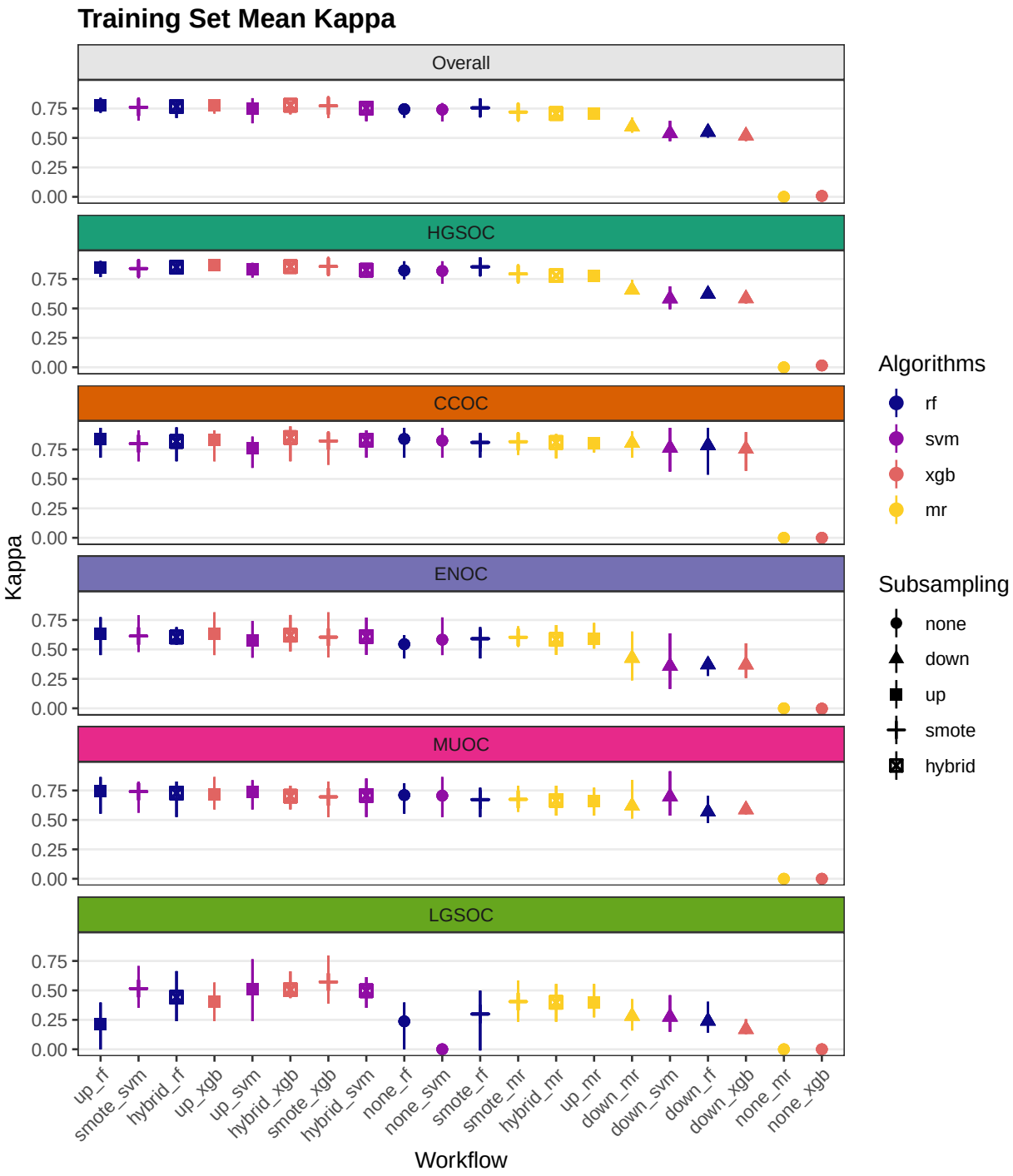


Figure 4.6: Training Set Mean Kappa

4.2 Rank Aggregation

Multi-step methods:

- **sequential**: sequential algorithm sequence of subsampling methods and algorithms used are:

  - HGSOC vs. non-HGSOC using upsampling and XGBoost
  - CCOC vs. non-CCOC using SMOTE subsampling and XGBoost
  - ENOC vs. non-ENOC using no subsampling and support vector machine
  - MUOC vs. LGSOC using SMOTE subsampling and random forest

- **two_step**: two-step algorithm sequence of subsampling methods and algorithms used are:
  - HGSOC vs. non-HGSOC using upsampling and XGBoost
  - CCOC vs. ENOC vs. MUOC vs. LGSOC using SMOTE subsampling and support vector machine

We conduct rank aggregation using a two-stage nested approach:

1. First we rank aggregate the per-class metrics for F1-score, balanced accuracy and kappa.
2. Then we take the aggregated lists from the three metrics and perform a final rank aggregation.
3. The top workflows from the final rank aggregation are used for gene optimization in the confirmation set

4.2.1 Across Classes

4.2.1.1 F1-Score

Table 4.7: F1-Score Rank Aggregation Summary

Workflow	Rank	HGSOC	CCOC	ENOC	MUOC	LGSOC
All	All	All	All	All	All	All
sequential	1	0.973	0.877	0.86	0.966	0.91
two_step	2	0.973	0.899	0.755	0.788	0.733
up_rf	3	0.972	0.849	0.652	0.757	0.362
smote_svm	4	0.969	0.811	0.638	0.752	0.524
hybrid_xgb	5	0.971	0.86	0.643	0.716	0.515
up_xgb	6	0.974	0.844	0.651	0.728	0.416
hybrid_svm	7	0.965	0.838	0.635	0.72	0.507
hybrid_rf	8	0.971	0.829	0.631	0.741	0.45
smote_xgb	9	0.972	0.833	0.628	0.71	0.579
smote_mr	10	0.956	0.827	0.629	0.691	0.418
up_svm	11	0.97	0.778	0.605	0.75	0.521
smote_rf	12	0.972	0.82	0.616	0.688	0.304
none_rf	13	0.968	0.848	0.568	0.723	0.4
hybrid_mr	14	0.952	0.821	0.613	0.682	0.412
up_mr	15	0.951	0.816	0.619	0.674	0.413
down_mr	16	0.913	0.817	0.47	0.638	0.302
down_svm	17	0.886	0.778	0.412	0.713	0.295
down_rf	18	0.901	0.799	0.419	0.589	0.262
down_xgb	19	0.886	0.771	0.42	0.608	0.194

4.2.1.2 Balanced Accuracy

Table 4.8: Balanced Accuracy Rank Aggregation Summary

Workflow	Rank	HGSOC	CCOC	ENOC	MUOC	LGSOC
All	All	All	All	All	All	All
sequential	1	0.926	0.907	0.856	0.943	0.943
hybrid_xgb	2	0.935	0.922	0.802	0.899	0.791
up_mr	3	0.93	0.917	0.817	0.881	0.844
smote_mr	4	0.932	0.912	0.813	0.889	0.844
smote_xgb	5	0.938	0.911	0.797	0.892	0.828
two_step	6	0.926	0.917	0.814	0.875	0.872
hybrid_mr	7	0.928	0.911	0.821	0.872	0.843
up_xgb	8	0.924	0.9	0.803	0.882	0.696
hybrid_svm	9	0.923	0.893	0.832	0.86	0.814
smote_svm	10	0.923	0.875	0.826	0.869	0.814
down_mr	11	0.906	0.909	0.776	0.87	0.874
hybrid_rf	12	0.923	0.898	0.806	0.894	0.689
up_svm	13	0.91	0.862	0.799	0.851	0.759
up_rf	14	0.905	0.894	0.808	0.878	0.59
smote_rf	15	0.911	0.872	0.797	0.865	0.595
down_rf	16	0.892	0.905	0.746	0.815	0.877
down_xgb	17	0.883	0.91	0.753	0.822	0.789
none_svm	18	0.882	0.879	0.778	0.868	0.5
none_rf	19	0.879	0.895	0.741	0.858	0.583
down_svm	20	0.879	0.873	0.757	0.906	0.906
none_mr	21	0.5	0.5	0.5	0.5	0.5
none_xgb	22	0.505	0.5	0.499	0.5	0.5

4.2.1.3 Kappa

Table 4.9: Kappa Rank Aggregation Summary

Workflow	Rank	HGSOC	CCOC	ENOC	MUOC	LGSOC
All	All	All	All	All	All	All
sequential	1	0.858	0.815	0.712	0.877	0.877
two_step	2	0.858	0.85	0.635	0.716	0.718
smote_svm	3	0.839	0.8	0.614	0.741	0.516
up_rf	4	0.85	0.84	0.63	0.746	0.213
up_xgb	5	0.865	0.835	0.63	0.714	0.408
hybrid_xgb	6	0.855	0.852	0.62	0.702	0.506
hybrid_svm	7	0.825	0.828	0.609	0.707	0.497
hybrid_rf	8	0.85	0.818	0.605	0.728	0.443
smote_xgb	9	0.856	0.822	0.604	0.695	0.571
up_svm	10	0.835	0.764	0.579	0.739	0.513
smote_mr	11	0.793	0.816	0.603	0.675	0.405
smote_rf	12	0.853	0.81	0.591	0.673	0.299
none_svm	13	0.818	0.825	0.583	0.706	0
none_rf	14	0.822	0.84	0.545	0.711	0.237
hybrid_mr	15	0.779	0.81	0.585	0.665	0.399
up_mr	16	0.778	0.803	0.592	0.657	0.399
down_mr	17	0.657	0.805	0.426	0.618	0.282
down_rf	18	0.621	0.785	0.37	0.568	0.24
down_xgb	19	0.584	0.754	0.37	0.588	0.17
down_svm	20	0.584	0.764	0.359	0.697	0.275
none_mr	21	0	0	0	0	0
none_xgb	22	0.016	0	-0.003	0	0

4.2.2 Across Metrics

Table 4.10: Rank Aggregation Comparison of Metrics Used in Training Set

Rank	F1	Balanced Accuracy	Kappa
1	sequential	sequential	sequential
2	two_step	hybrid_xgb	two_step
3	up_rf	up_mr	smote_svm
4	smote_svm	smote_mr	up_rf
5	hybrid_xgb	smote_xgb	up_xgb
6	up_xgb	two_step	hybrid_xgb
7	hybrid_svm	hybrid_mr	hybrid_svm
8	hybrid_rf	up_xgb	hybrid_rf
9	smote_xgb	hybrid_svm	smote_xgb
10	smote_mr	smote_svm	up_svm
11	up_svm	down_mr	smote_mr
12	smote_rf	hybrid_rf	smote_rf
13	none_rf	up_svm	none_svm
14	hybrid_mr	up_rf	none_rf
15	up_mr	smote_rf	hybrid_mr
16	down_mr	down_rf	up_mr
17	down_svm	down_xgb	down_mr
18	down_rf	none_svm	down_rf
19	down_xgb	none_rf	down_xgb
20	NA	down_svm	down_svm
21	NA	none_mr	none_mr
22	NA	none_xgb	none_xgb

Table 4.11: Top 5 Workflows from Final Rank Aggregation

Rank	Workflow
1	sequential
2	two_step
3	up_rf
4	smote_svm
5	hybrid_xgb

4.2.3 Top Workflows

We look at the per-class evaluation metrics of the top 5 workflows.

Table 4.12: Top Workflow Per-Class Evaluation Metrics

		Histotypes				
Metric	Workflow	HGSOC	CCOC	ENOC	MUOC	LGSOC
Accuracy	Sequential	0.96 (0.95, 0.96)	0.92 (0.88, 0.96)	0.86 (0.77, 0.91)	0.95 (0.88, 1)	0.95 (0.88, 1)
	2-STEP	0.96 (0.95, 0.96)	0.93 (0.9, 0.96)	0.84 (0.73, 0.9)	0.89 (0.81, 0.96)	0.97 (0.94, 1)
	Up-RF	0.95 (0.92, 0.98)	0.98 (0.97, 0.99)	0.96 (0.95, 0.97)	0.98 (0.96, 0.99)	0.98 (0.97, 0.99)
	SMOTE-SVM	0.95 (0.94, 0.97)	0.98 (0.97, 0.99)	0.95 (0.94, 0.96)	0.98 (0.96, 0.98)	0.98 (0.97, 0.98)
	Hybrid-XGB	0.95 (0.94, 0.97)	0.98 (0.97, 0.99)	0.96 (0.94, 0.97)	0.97 (0.96, 0.98)	0.98 (0.97, 0.99)
Sensitivity	Sequential	0.98 (0.97, 0.99)	0.88 (0.81, 1)	0.86 (0.81, 0.94)	0.97 (0.91, 1)	0.92 (0.6, 1)
	2-STEP	0.98 (0.97, 0.99)	0.87 (0.82, 0.93)	0.73 (0.56, 0.82)	0.84 (0.75, 0.92)	0.77 (0, 1)
	Up-RF	0.99 (0.98, 1)	0.79 (0.57, 0.88)	0.64 (0.44, 0.81)	0.77 (0.46, 0.91)	0.18 (0, 0.33)
	SMOTE-SVM	0.97 (0.96, 0.98)	0.76 (0.57, 0.85)	0.68 (0.5, 0.86)	0.75 (0.54, 0.91)	0.64 (0.5, 0.75)
	Hybrid-XGB	0.97 (0.96, 0.98)	0.85 (0.57, 0.95)	0.62 (0.44, 0.77)	0.82 (0.54, 0.91)	0.59 (0.38, 0.75)
Specificity	Sequential	0.88 (0.85, 0.9)	0.94 (0.85, 1)	0.85 (0.73, 0.88)	0.92 (0.6, 1)	0.97 (0.91, 1)
	2-STEP	0.88 (0.85, 0.9)	0.97 (0.94, 1)	0.89 (0.83, 0.94)	0.91 (0.83, 0.97)	0.98 (0.95, 1)
	Up-RF	0.82 (0.74, 0.89)	1 (0.99, 1)	0.98 (0.97, 0.99)	0.99 (0.98, 1)	1 (0.99, 1)
	SMOTE-SVM	0.88 (0.8, 0.92)	0.99 (0.99, 1)	0.97 (0.96, 0.98)	0.99 (0.98, 1)	0.99 (0.98, 1)
	Hybrid-XGB	0.9 (0.85, 0.95)	0.99 (0.99, 1)	0.98 (0.97, 0.99)	0.98 (0.97, 0.99)	0.99 (0.98, 0.99)
F1-Score	Sequential	0.97 (0.97, 0.98)	0.88 (0.81, 0.94)	0.86 (0.79, 0.91)	0.97 (0.92, 1)	0.91 (0.75, 1)
	2-STEP	0.97 (0.97, 0.98)	0.9 (0.85, 0.93)	0.75 (0.61, 0.85)	0.79 (0.67, 0.92)	0.73 (0, 1)
	Up-RF	0.97 (0.95, 0.99)	0.85 (0.7, 0.94)	0.65 (0.47, 0.79)	0.76 (0.57, 0.87)	0.36 (0.29, 0.4)
	SMOTE-SVM	0.97 (0.96, 0.98)	0.81 (0.67, 0.92)	0.64 (0.5, 0.81)	0.75 (0.58, 0.83)	0.52 (0.36, 0.71)
	Hybrid-XGB	0.97 (0.96, 0.98)	0.86 (0.67, 0.95)	0.64 (0.5, 0.81)	0.72 (0.61, 0.8)	0.51 (0.44, 0.67)
Balanced Accuracy	Sequential	0.93 (0.91, 0.94)	0.91 (0.86, 0.95)	0.86 (0.77, 0.91)	0.94 (0.8, 1)	0.94 (0.8, 1)
	2-STEP	0.93 (0.91, 0.94)	0.92 (0.88, 0.95)	0.81 (0.69, 0.88)	0.87 (0.79, 0.95)	0.87 (0.48, 1)
	Up-RF	0.9 (0.86, 0.94)	0.89 (0.78, 0.94)	0.81 (0.71, 0.9)	0.88 (0.73, 0.95)	0.59 (0.5, 0.66)
	SMOTE-SVM	0.92 (0.89, 0.95)	0.88 (0.78, 0.92)	0.83 (0.74, 0.92)	0.87 (0.76, 0.95)	0.81 (0.75, 0.87)
	Hybrid-XGB	0.94 (0.91, 0.96)	0.92 (0.78, 0.97)	0.8 (0.72, 0.88)	0.9 (0.76, 0.95)	0.79 (0.68, 0.87)
Kappa	Sequential	0.86 (0.83, 0.88)	0.82 (0.72, 0.91)	0.71 (0.55, 0.82)	0.88 (0.68, 1)	0.88 (0.68, 1)
	2-STEP	0.86 (0.83, 0.88)	0.85 (0.77, 0.9)	0.63 (0.4, 0.77)	0.72 (0.54, 0.9)	0.72 (-0.03, 1)
	Up-RF	0.85 (0.77, 0.9)	0.84 (0.68, 0.93)	0.63 (0.45, 0.77)	0.75 (0.55, 0.86)	0.21 (0, 0.4)
	SMOTE-SVM	0.84 (0.78, 0.88)	0.8 (0.65, 0.91)	0.61 (0.48, 0.79)	0.74 (0.56, 0.83)	0.52 (0.35, 0.71)
	Hybrid-XGB	0.85 (0.83, 0.89)	0.85 (0.65, 0.95)	0.62 (0.48, 0.79)	0.7 (0.59, 0.79)	0.51 (0.44, 0.66)

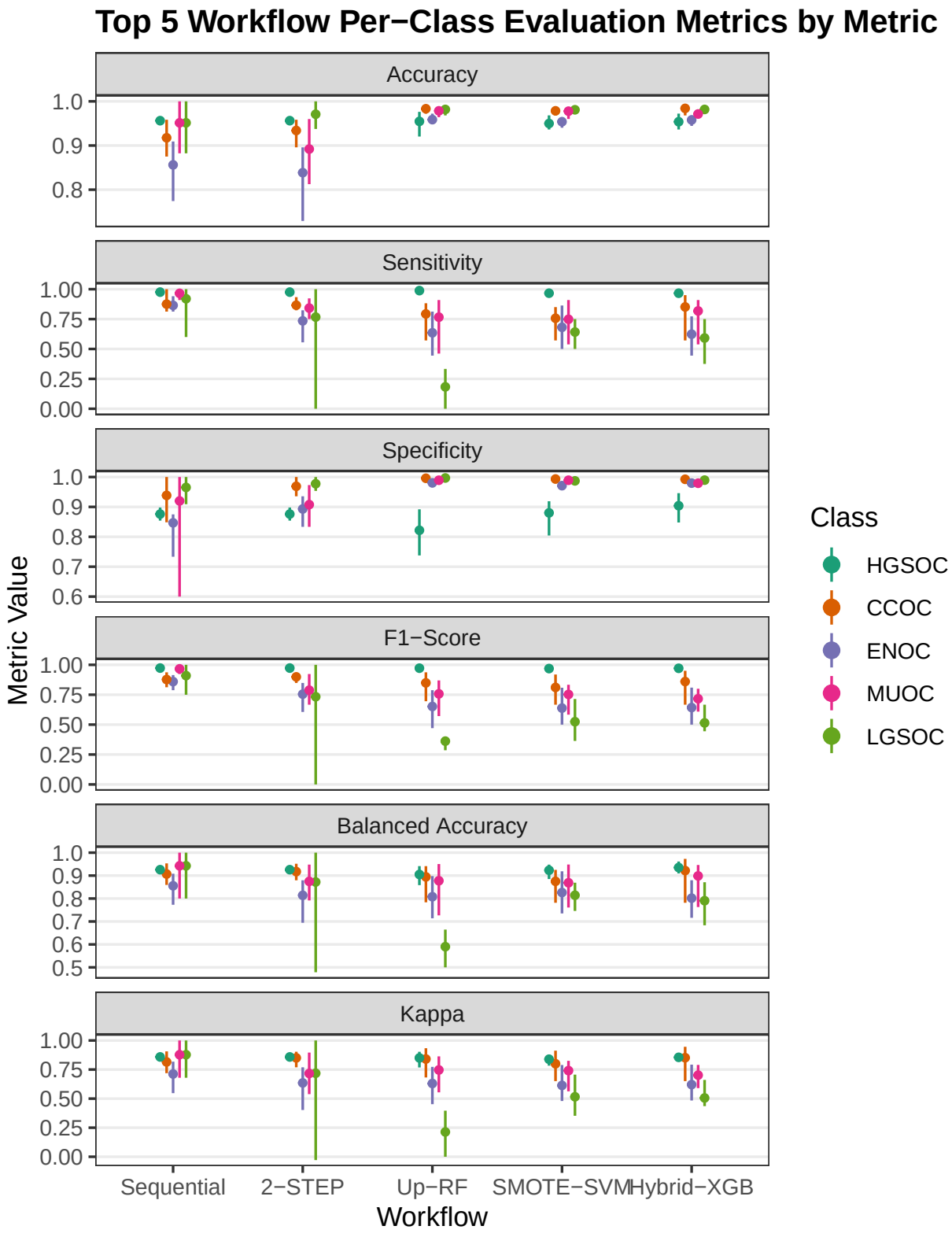


Figure 4.7: Top 5 Workflow Per-Class Evaluation Metrics by Metric

Table 4.13: Top Workflow Per-Class Evaluation Metrics and Ranks

Workflow	Rank	HGSOC	CCOC	ENOC	MUOC	LGSOC
F1-Score						
Sequential	1	0.973	0.877	0.860	0.966	0.910
2-STEP	2	0.973	0.899	0.755	0.788	0.733
Up-RF	3	0.972	0.849	0.652	0.757	0.362
SMOTE-SVM	4	0.969	0.811	0.638	0.752	0.524
Hybrid-XGB	5	0.971	0.860	0.643	0.716	0.515
Balanced Accuracy						
Sequential	1	0.926	0.907	0.856	0.943	0.943
Hybrid-XGB	2	0.935	0.922	0.802	0.899	0.791
2-STEP	6	0.926	0.917	0.814	0.875	0.872
SMOTE-SVM	10	0.923	0.875	0.826	0.869	0.814
Up-RF	14	0.905	0.894	0.808	0.878	0.590
Kappa						
Sequential	1	0.858	0.815	0.712	0.877	0.877
2-STEP	2	0.858	0.850	0.635	0.716	0.718
SMOTE-SVM	3	0.839	0.800	0.614	0.741	0.516
Up-RF	4	0.850	0.840	0.630	0.746	0.213
Hybrid-XGB	6	0.855	0.852	0.620	0.702	0.506

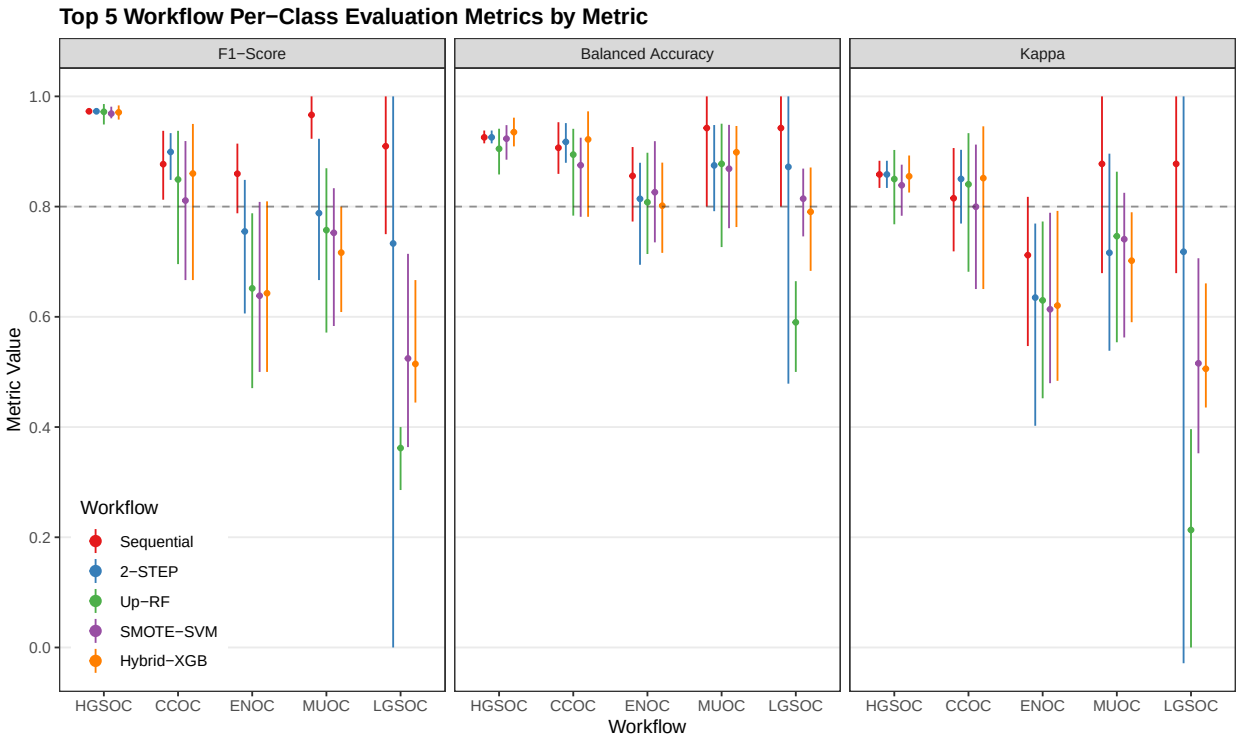


Figure 4.8: Top 5 Workflow Per-Class Evaluation Metrics by Metric

Misclassified cases from a previous step of the sequence of classifiers are not included in subsequent steps of the training set CV folds. Thus, we cannot piece together the test set predictions from the sequential and two-step algorithms to obtain overall metrics.

4.3 Confirmation Set

Now we'd like to see how our best five workflows perform in the confirmation set. The class-specific F1-scores will be used. The top performing method will be selected for gene optimization.

Table 4.14: Evaluation Metrics on Confirmation Set Models

			Histotypes				
Method	Metric	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
Sequential	Accuracy	0.829	0.866	0.969	0.883	0.966	0.974
	Sensitivity	0.584	0.953	0.861	0.467	0.556	0.083
	Specificity	0.924	0.697	0.982	0.966	0.984	0.990
	F1-Score	0.604	0.904	0.861	0.571	0.577	0.105
	Balanced Accuracy	0.754	0.825	0.922	0.717	0.770	0.537
	Kappa	0.646	0.685	0.844	0.508	0.559	0.093
2-STEP	Accuracy	0.838	0.866	0.970	0.891	0.977	0.972
	Sensitivity	0.613	0.953	0.875	0.486	0.667	0.083
	Specificity	0.926	0.697	0.982	0.972	0.990	0.989
	F1-Score	0.635	0.904	0.869	0.598	0.706	0.100
	Balanced Accuracy	0.769	0.825	0.929	0.729	0.828	0.536
	Kappa	0.666	0.685	0.852	0.538	0.694	0.086
Up-RF	Accuracy	0.835	0.857	0.975	0.883	0.974	0.981
	Sensitivity	0.613	0.972	0.875	0.383	0.667	0.167
	Specificity	0.918	0.633	0.988	0.983	0.987	0.997
	F1-Score	0.648	0.900	0.887	0.522	0.679	0.250
	Balanced Accuracy	0.765	0.802	0.931	0.683	0.827	0.582
	Kappa	0.646	0.654	0.873	0.466	0.665	0.243
SMOTE-SVM	Accuracy	0.827	0.866	0.958	0.888	0.972	0.970
	Sensitivity	0.650	0.939	0.861	0.477	0.556	0.417
	Specificity	0.927	0.725	0.970	0.970	0.990	0.981
	F1-Score	0.656	0.902	0.821	0.586	0.625	0.345
	Balanced Accuracy	0.788	0.832	0.916	0.723	0.773	0.699
	Kappa	0.651	0.690	0.797	0.525	0.611	0.330
Hybrid-XGB	Accuracy	0.830	0.869	0.961	0.893	0.967	0.970
	Sensitivity	0.662	0.943	0.861	0.458	0.630	0.417
	Specificity	0.928	0.725	0.974	0.979	0.982	0.981
	F1-Score	0.657	0.905	0.832	0.587	0.618	0.345
	Balanced Accuracy	0.795	0.834	0.917	0.719	0.806	0.699
	Kappa	0.657	0.696	0.810	0.531	0.601	0.330

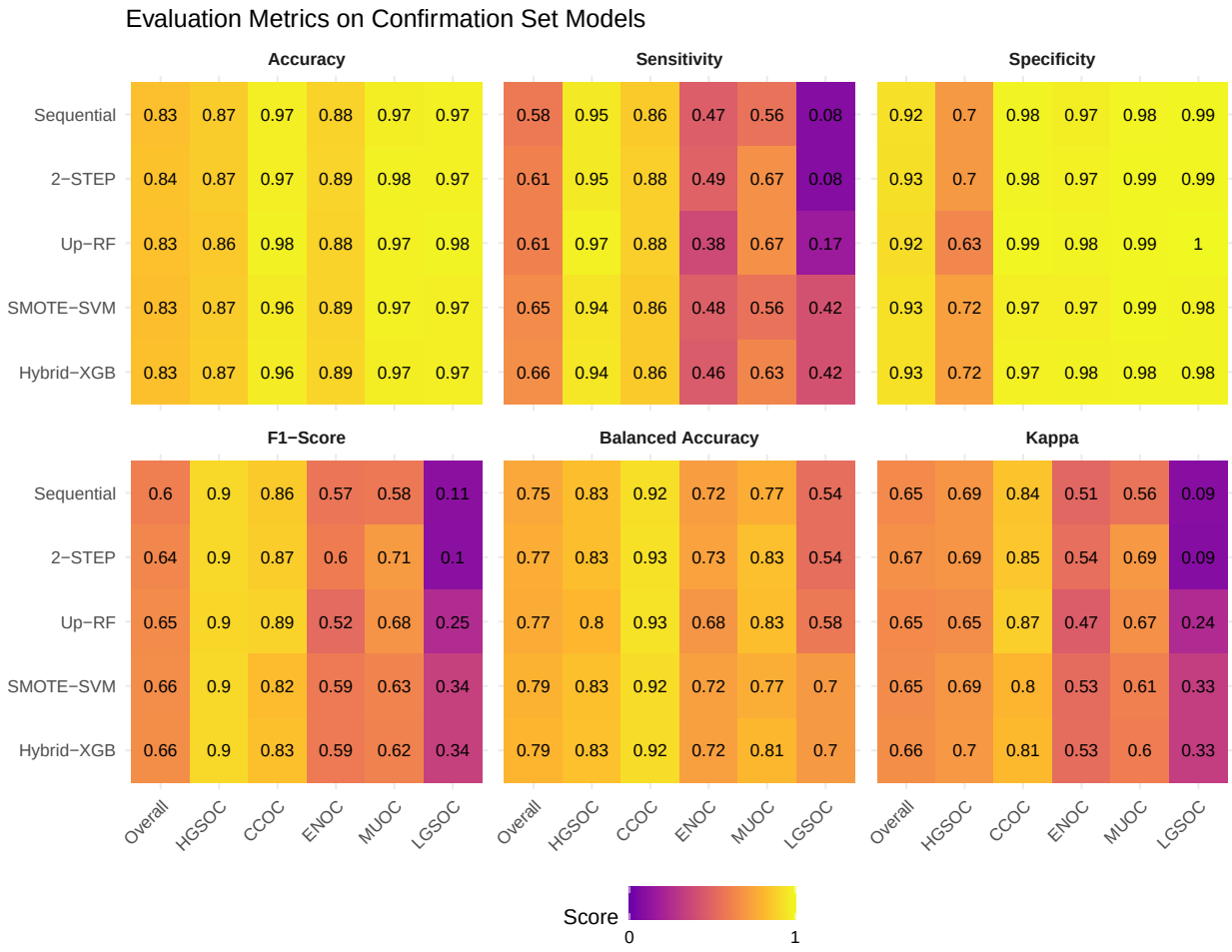


Figure 4.9: Evaluation Metrics on Confirmation Set Models

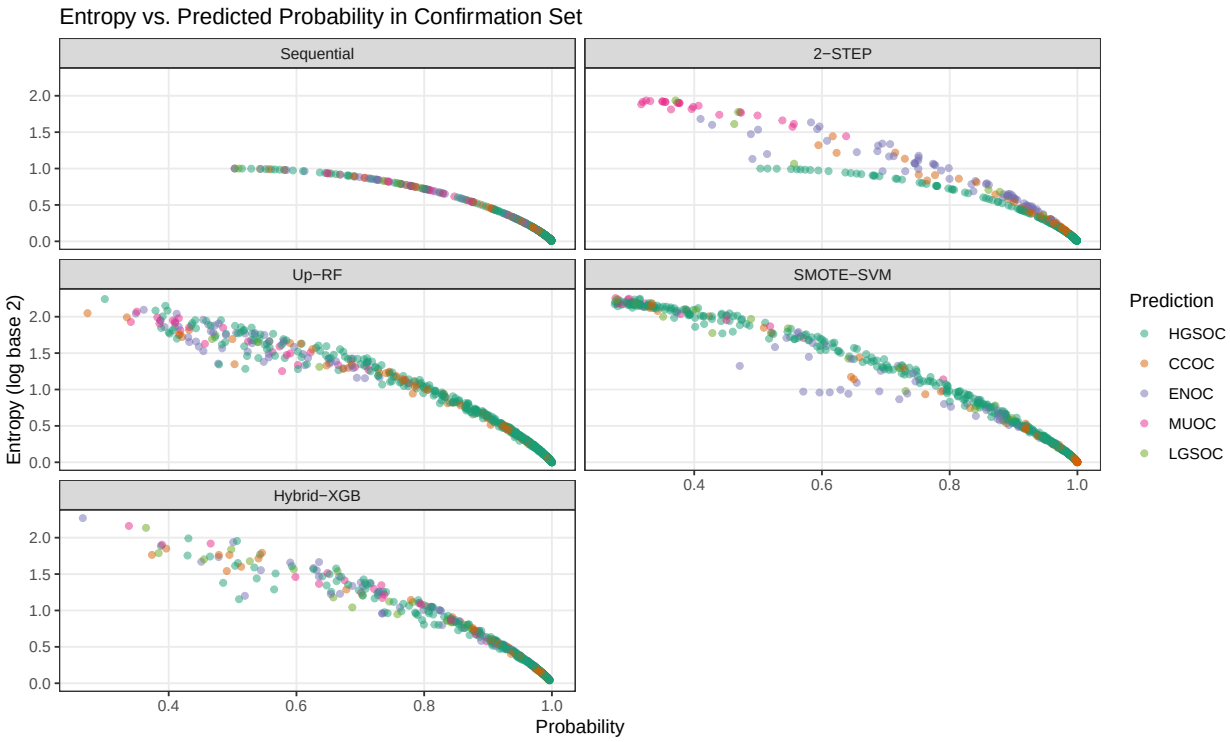


Figure 4.10: Entropy vs. Predicted Probability in Confirmation Set

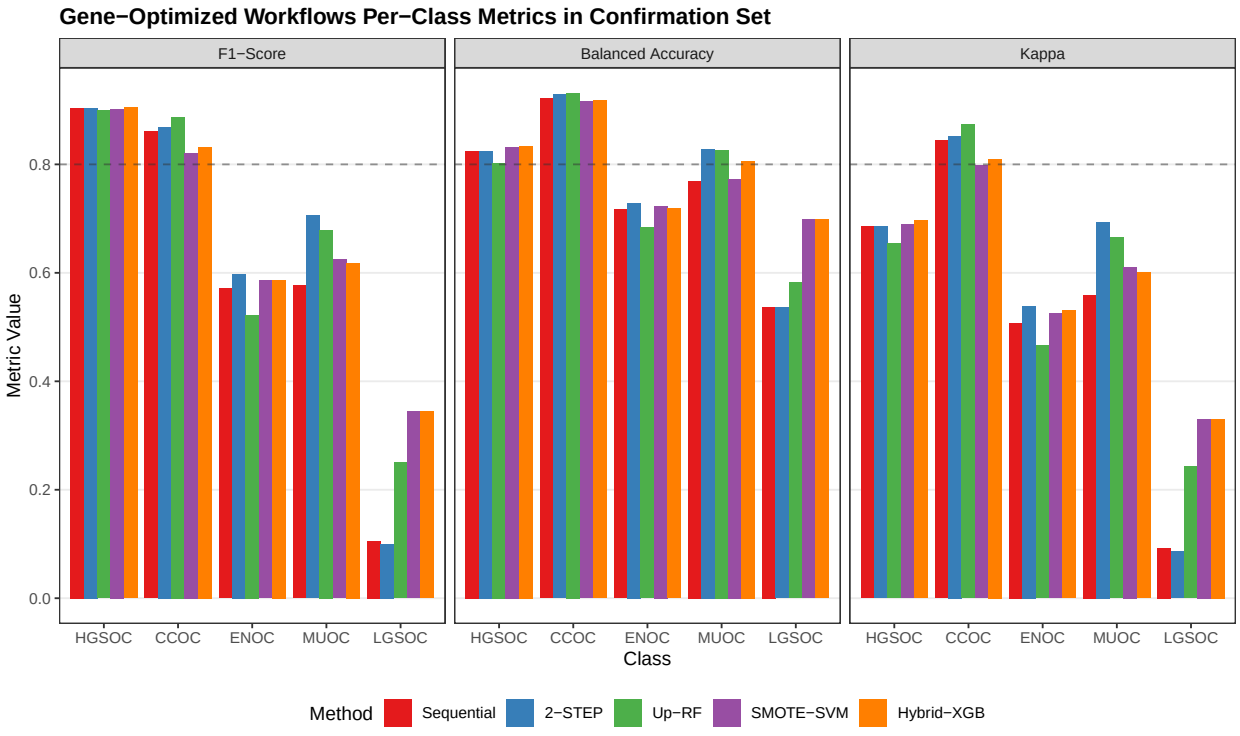


Figure 4.11: Gene Optimized Workflows Per-Class Metrics in Confirmation Set

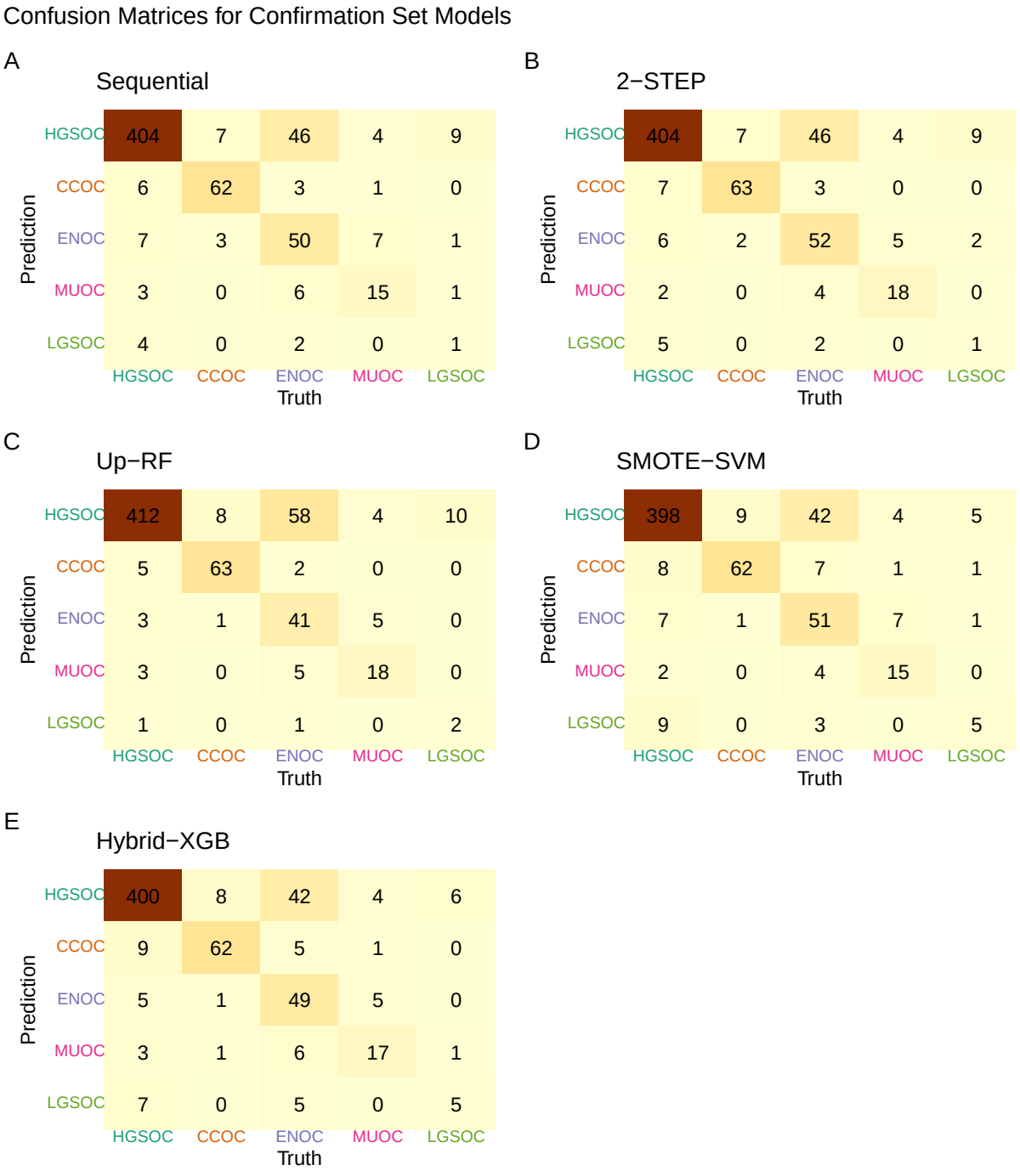


Figure 4.12: Confusion Matrices for Confirmation Set Models

4.3.1 Sequential

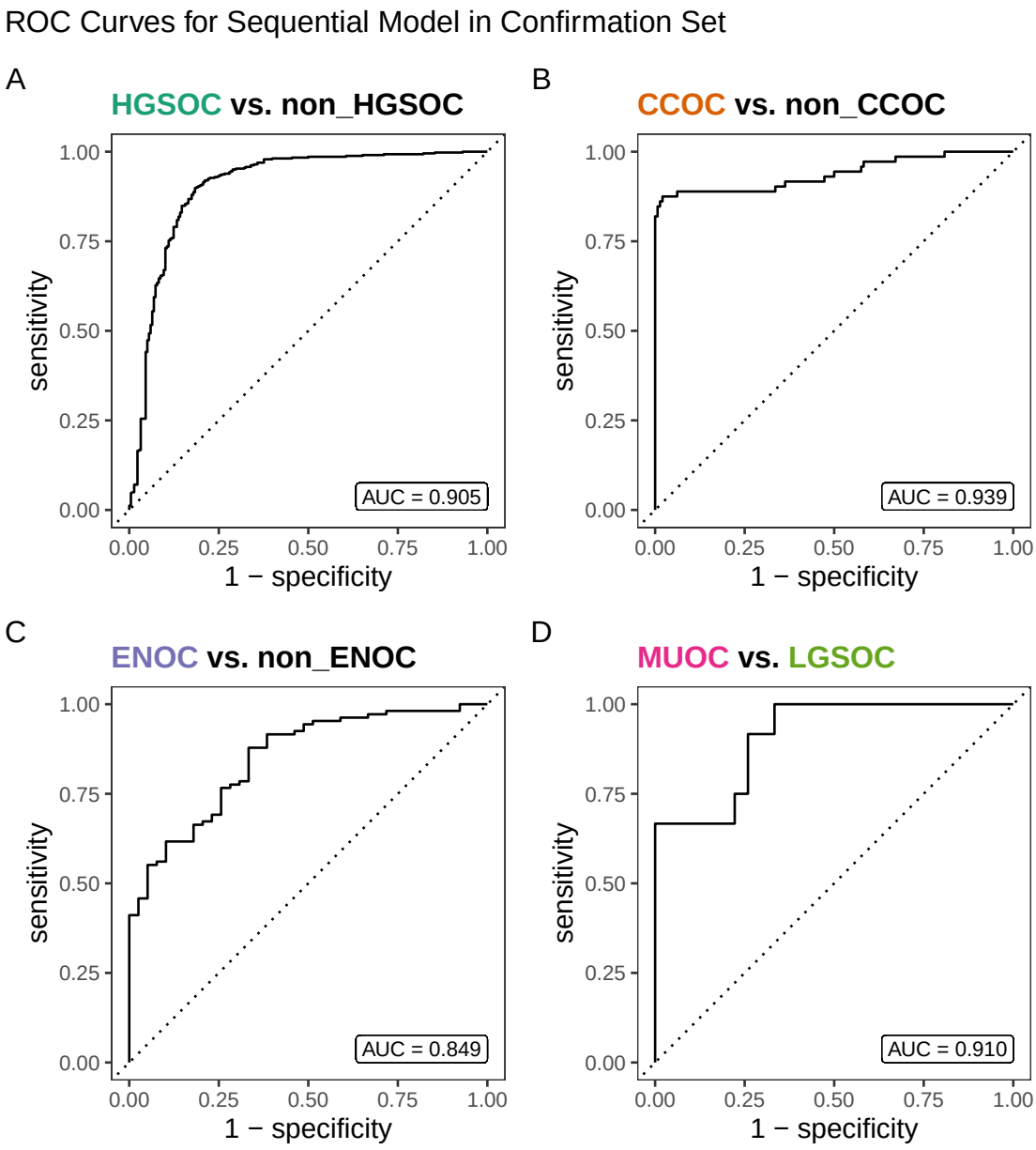


Figure 4.13: ROC Curves for Sequential Model in Confirmation Set

4.3.2 2-STEP

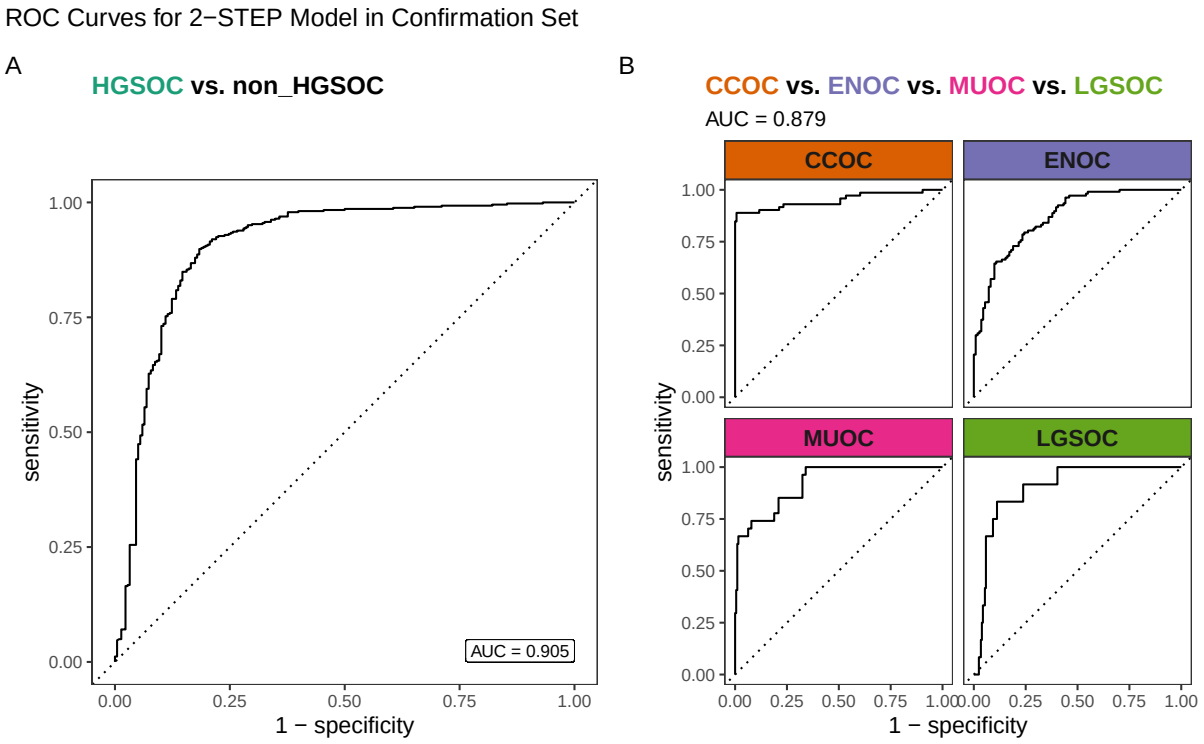


Figure 4.14: ROC Curves for 2-STEP Model in Confirmation Set

4.3.3 Up-RF

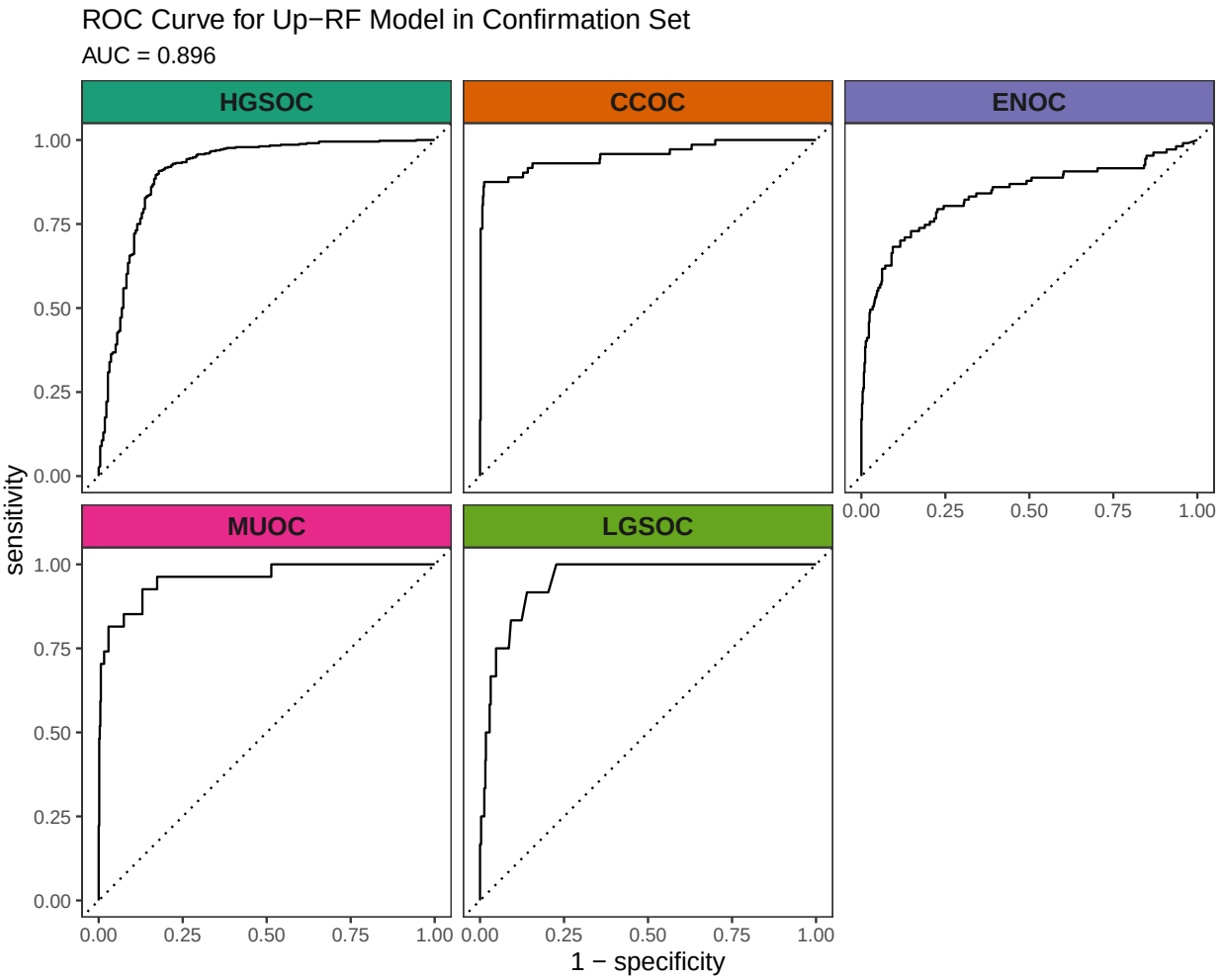


Figure 4.15: ROC Curve for Up-RF Model in Confirmation Set

4.3.4 SMOTE-SVM

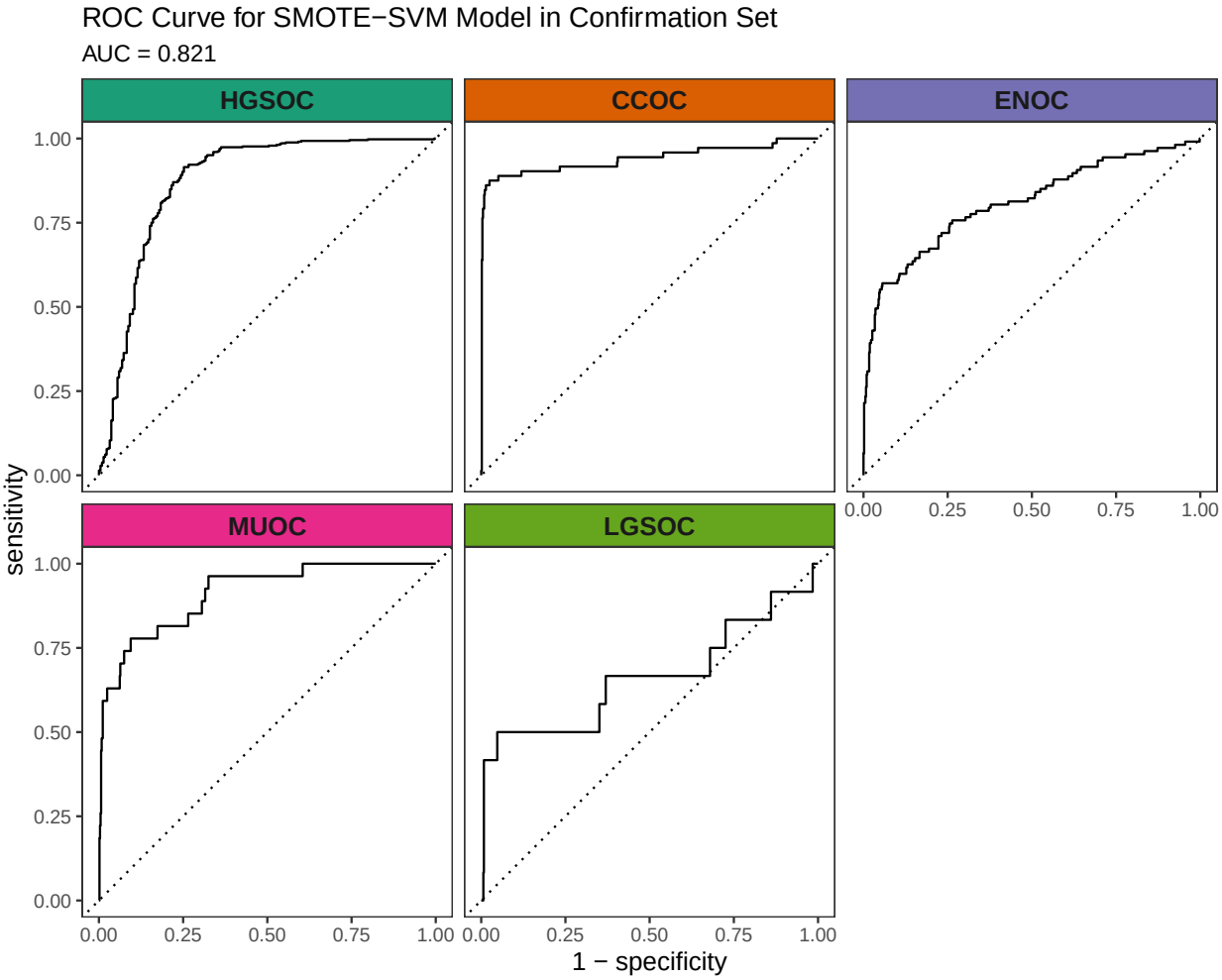


Figure 4.16: ROC Curve for SMOTE-SVM Model in Confirmation Set

4.3.5 Hybrid-XGB

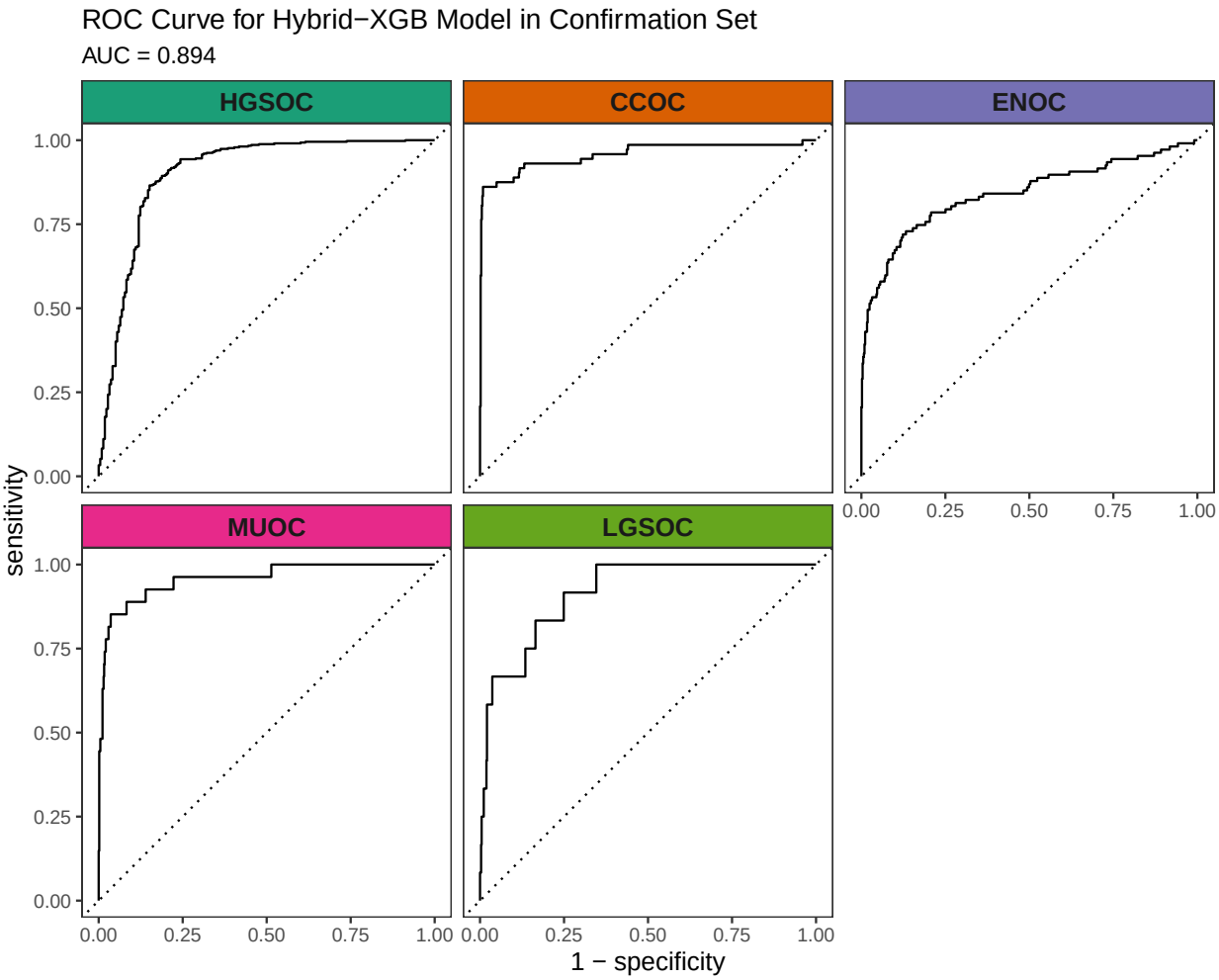


Figure 4.17: ROC Curve for Hybrid-XGB Model in Confirmation Set

4.4 Gene Optimization

From Figure 4.9, we see that both Hybrid-XGB and SMOTE-SVM have the highest overall F1-Score and are relatively better at predicting the rarest histotype LGSOC (sensitivity = 0.42). Thus we choose both of these two workflows for gene optimization in the confirmation set. The optimal number of genes is determined by the highest average F1-Score across classes, including the overall metric in the average. We use an F1-score that is averaged across cross-validation folds (5) and class groups (6: Overall, HGSOC, CCOC, ENOC, MUOC, LGSOC) to compare performance between different number of genes selected.

4.4.1 Hybrid-XGB

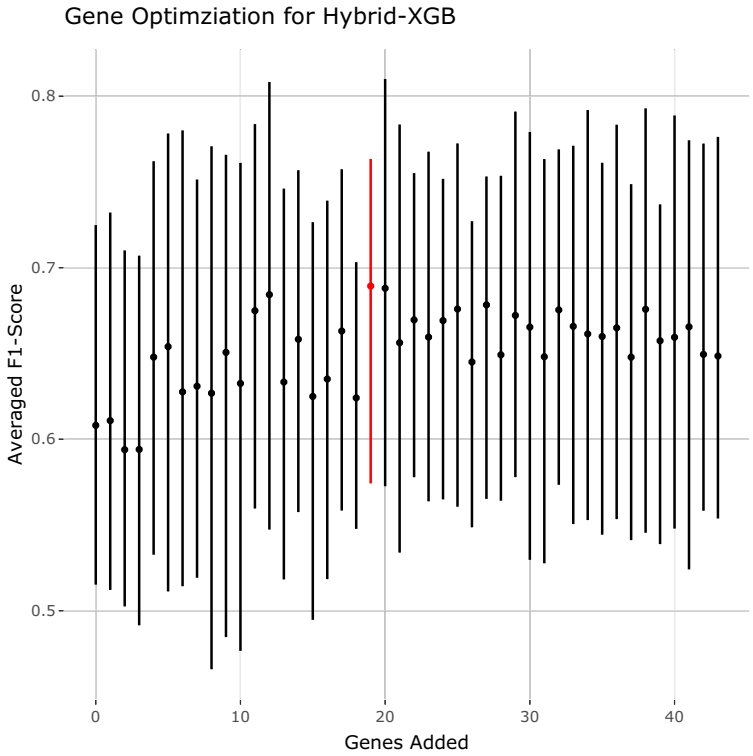


Figure 4.18: Gene Optimization for Hybrid-XGB Classifier using Averaged F1-Score

In the Hybrid-XGB classifier, the optimal number of genes is highlighted in red in Figure 4.18. Hence the optimal number of total genes used will be n=28+19=47.

The gene profile of the optimal set of genes used is displayed in Table 4.15. Base genes in the PrOTYPE and SPOT sets are annotated with green circles, and the added genes are annotated with yellow circles. The added genes are: HNF1B, TFF1, TPX2, WT1, IGFBP1, LGALS4, TFF3, KLK7, CYP2C18, GPR64, CAPN2, MET, GCNT3, GAD1, SLC3A1, EGFL6, C1orf173, DKK4 and C10orf116. Unused genes are annotated with red crosses.

Table 4.15: Gene Profile of Optimal Set in Hybrid-XGB Workflow

Set	Genes	PrOTYPE	SPOT	Optimal Set	Candidate Rank
	COL11A1	v		(*)	
	CD74	v		(*)	
	CD2	v		(*)	
	TIMP3	v		(*)	
	LUM	v		(*)	
	CYTIP	v		(*)	
	COL3A1	v		(*)	

	THBS2	v		(*)	
	TCF7L1	v	v	(*)	
	HMGA2	v		(*)	
	FN1	v		(*)	
	POSTN	v		(*)	
	COL1A2	v		(*)	
	COL5A2	v		(*)	
	PDZK1IP1	v		(*)	
	FBN1	v		(*)	
	HIF1A		v	(*)	
Base	CXCL10		v	(*)	
	DUSP4		v	(*)	
	SOX17		v	(*)	
	MITF		v	(*)	
	CDKN3		v	(*)	
	BRCA2		v	(*)	
	CEACAM5		v	(*)	
	ANXA4		v	(*)	
	SERPINE1		v	(*)	
	CRABP2		v	(*)	
	DNAJC9		v	(*)	
	HNF1B			(*)	1
	TFF1			(*)	2
	TPX2			(*)	3
	WT1			(*)	4
	IGFBP1			(*)	5
	LGALS4			(*)	6
	TFF3			(*)	7
	KLK7			(*)	8
	CYP2C18			(*)	9
	GPR64			(*)	10
	CAPN2			(*)	11
	MET			(*)	12
	GCNT3			(*)	13

Candidates	GAD1	(*)	14
	SLC3A1	(*)	15
	EGFL6	(*)	16
	C1orf173	(*)	17
	DKK4	(*)	18
	C10orf116	(*)	19
	FUT3	(x)	20
	PBX1	(x)	21
	MUC5B	(x)	22
	KGFLP2	(x)	23
	IGKC	(x)	24
	IL6	(x)	25
	CPNE8	(x)	26
	CYP4B1	(x)	27
	TP53	(x)	28
	PAX8	(x)	29
	SERPINA5	(x)	30
	SENP8	(x)	31
	BRCA1	(x)	32
	STC1	(x)	33
	SEMA6A	(x)	34
	TSPAN8	(x)	35
	LIN28B	(x)	36
	EPAS1	(x)	37
	ATP5G3	(x)	38
	IGJ	(x)	39
	SCGB1D2	(x)	40
	BCL2	(x)	41
	ADCYAP1R1	(x)	42
	MAP1LC3A	(x)	43

4.4.2 SMOTE-SVM

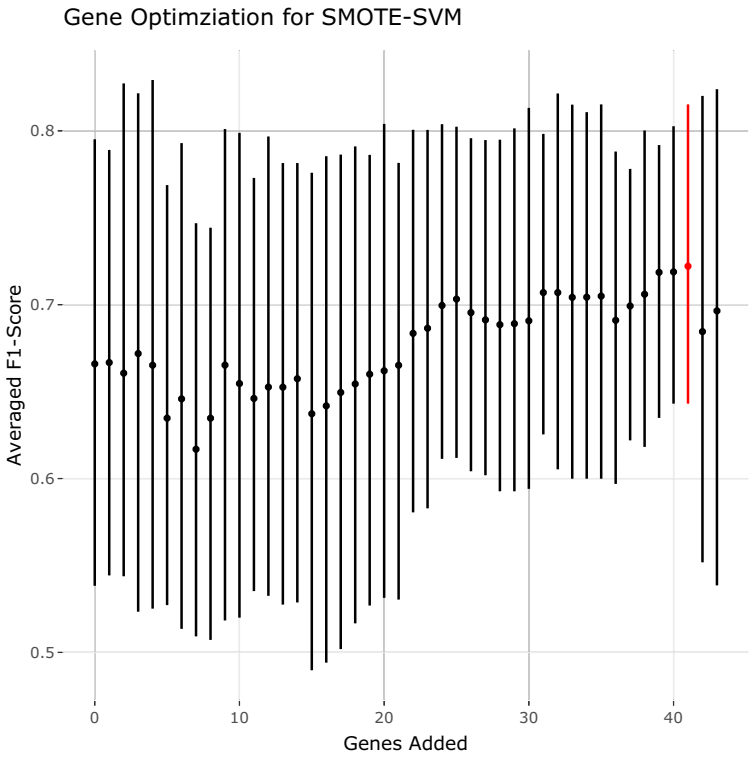


Figure 4.19: Gene Optimization for SMOTE-SVM Classifier using Averaged F1-Score

In the SMOTE-svm classifier, the optimal number of genes is achieved at the highest averaged F1-score with 41 genes added, highlighted in red in Figure 4.19. Hence the optimal number of total genes used will be n=28+41=69.

The gene profile of the optimal set of genes used is displayed in Table 4.16. Base genes in the PrOTYPE and SPOT sets are annotated with green circles, and the added genes are annotated with yellow circles. The added genes are: EGFL6, IGJ, IGKC, TP53, DKK4, MUC5B, SLC3A1, MAP1LC3A, IGFBP1, CPNE8, SERPINA5, SCGB1D2, STC1, EPAS1, BRCA1, KGFLP2, SENP8, BCL2, PBX1, KLK7, C10orf116, LIN28B, LGALS4, ADCYAP1R1, IL6, ZBED1, WT1, TFF1, GCNT3, HNF1B, TFF3, CYP4B1, CYP2C18, TSPAN8, FUT3, MET, ATP5G3, SEMA6A, GPR64, PAX8 and C1orf173. Unused genes are annotated with red crosses.

Table 4.16: Gene Profile of Optimal Set in SMOTE-SVM Workflow

Set	Genes	PrOTYPE	SPOT	Optimal Set	Candidate Rank
	COL11A1	v		(*)	
	CD74	v		(*)	
	CD2	v		(*)	
	TIMP3	v		(*)	
	LUM	v		(*)	

	CYTIP	v		(*)	
	COL3A1	v		(*)	
	THBS2	v		(*)	
	TCF7L1	v	v	(*)	
	HMGA2	v		(*)	
	FN1	v		(*)	
	POSTN	v		(*)	
	COL1A2	v		(*)	
	COL5A2	v		(*)	
	PDZK1IP1	v		(*)	
	FBN1	v		(*)	
	HIF1A		v	(*)	
Base	CXCL10		v	(*)	
	DUSP4		v	(*)	
	SOX17		v	(*)	
	MITF		v	(*)	
	CDKN3		v	(*)	
	BRCA2		v	(*)	
	CEACAM5		v	(*)	
	ANXA4		v	(*)	
	SERPINE1		v	(*)	
	CRABP2		v	(*)	
	DNAJC9		v	(*)	
	EGFL6			(*)	1
	IGJ			(*)	2
	IGKC			(*)	3
	TP53			(*)	4
	DKK4			(*)	5
	MUC5B			(*)	6
	SLC3A1			(*)	7
	MAP1LC3A			(*)	8
	IGFBP1			(*)	9
	CPNE8			(*)	10
	SERPINA5			(*)	11

Candidates	SCGB1D2	(*)	12
	STC1	(*)	13
	EPAS1	(*)	14
	BRCA1	(*)	15
	KGFLP2	(*)	16
	SENP8	(*)	17
	BCL2	(*)	18
	PBX1	(*)	19
	KLK7	(*)	20
	C10orf116	(*)	21
	LIN28B	(*)	22
	LGALS4	(*)	23
	ADCYAP1R1	(*)	24
	IL6	(*)	25
	ZBED1	(*)	26
	WT1	(*)	27
	TFF1	(*)	28
	GCNT3	(*)	29
	HNF1B	(*)	30
	TFF3	(*)	31
	CYP4B1	(*)	32
	CYP2C18	(*)	33
	TSPAN8	(*)	34
	FUT3	(*)	35
	MET	(*)	36
	ATP5G3	(*)	37
	SEMA6A	(*)	38
	GPR64	(*)	39
	PAX8	(*)	40
	C1orf173	(*)	41
	GAD1	(x)	42
	CAPN2	(x)	43

4.4.3 Gene List Comparisons in Confirmation Set

We train the Hybrid-XGB and SMOTE-SVM workflows using the base and optimal gene lists in the training set. The models are evaluated on the confirmation set. Overall and per-class results are shown in Table 4.17. The gene lists are:

1. Base (n=28): among the overlapping genes, the base set from the PrOTYPE and SPOT lists
2. Optimal (n=47, 69): among the overlapping genes, the base set plus the additional number of genes that result in the optimal value for a selected evaluation metric, as assessed in Figure 4.18 and Figure 4.19

Table 4.17: Model Comparisons using Different Gene Lists in Confirmation Set

			Histotypes				
Method	Metric	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
Hybrid-XGB, Optimal	Accuracy	0.836	0.872	0.966	0.893	0.966	0.977
	Sensitivity	0.694	0.946	0.861	0.458	0.704	0.500
	Specificity	0.930	0.729	0.979	0.979	0.977	0.986
	F1-Score	0.684	0.907	0.849	0.587	0.633	0.444
	Balanced Accuracy	0.812	0.838	0.920	0.719	0.840	0.743
	Kappa	0.670	0.703	0.830	0.531	0.616	0.433
Hybrid-XGB, Base	Accuracy	0.810	0.858	0.952	0.872	0.970	0.967
	Sensitivity	0.613	0.929	0.875	0.393	0.704	0.167
	Specificity	0.923	0.720	0.961	0.968	0.982	0.983
	F1-Score	0.606	0.896	0.803	0.506	0.667	0.160
	Balanced Accuracy	0.768	0.825	0.918	0.680	0.843	0.575
	Kappa	0.620	0.673	0.775	0.440	0.651	0.143
SMOTE-SVM, Optimal	Accuracy	0.821	0.860	0.958	0.879	0.974	0.972
	Sensitivity	0.666	0.932	0.847	0.458	0.593	0.500
	Specificity	0.925	0.720	0.972	0.963	0.990	0.981
	F1-Score	0.665	0.898	0.819	0.557	0.653	0.400
	Balanced Accuracy	0.796	0.826	0.910	0.710	0.791	0.740
	Kappa	0.640	0.676	0.795	0.490	0.639	0.386
SMOTE-SVM, Base	Accuracy	0.815	0.860	0.952	0.874	0.970	0.974
	Sensitivity	0.672	0.903	0.889	0.514	0.556	0.500
	Specificity	0.930	0.775	0.960	0.946	0.989	0.983
	F1-Score	0.660	0.895	0.805	0.576	0.612	0.414
	Balanced Accuracy	0.801	0.839	0.924	0.730	0.772	0.741
	Kappa	0.641	0.685	0.778	0.503	0.597	0.401

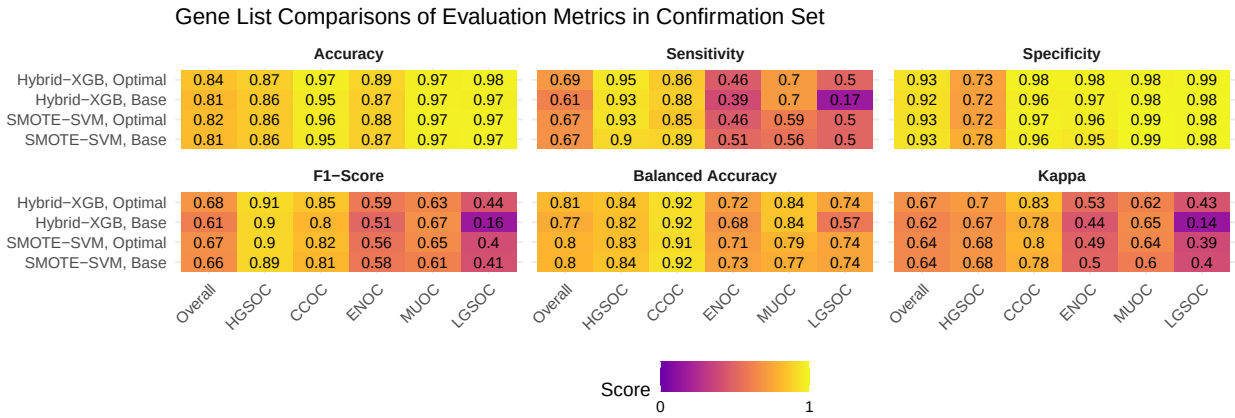


Figure 4.20: Gene List Comparisons of Evaluation Metrics in Confirmation Set

4.5 Validation Set

From the results in Table 4.17 and Figure 4.20, we see that the Hybrid-XGB, optimal workflow as the highest overall F1-Score and balanced accuracy. Thus, we choose the Hybrid-XGB model trained on the training set for the all overlap, optimal, and base gene lists, and evaluate performance in the validation set.

4.5.1 Evaluation Metrics

Table 4.18: Evaluation Metrics on Training Set Models in Validation Set

			Histotypes				
Method	Metric	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
Hybrid-XGB, All Overlap	Accuracy	0.889	0.905	0.969	0.955	0.978	0.972
	Sensitivity	0.799	0.913	1.000	0.682	0.870	0.533
	Specificity	0.958	0.877	0.966	0.985	0.980	0.980
	F1-Score	0.715	0.938	0.831	0.750	0.667	0.390
	Balanced Accuracy	0.879	0.895	0.983	0.833	0.925	0.756
	Kappa	0.725	0.739	0.815	0.726	0.656	0.377
Hybrid-XGB, Optimal	Accuracy	0.876	0.894	0.966	0.949	0.974	0.969
	Sensitivity	0.789	0.904	0.971	0.625	0.913	0.533
	Specificity	0.952	0.856	0.966	0.984	0.976	0.976
	F1-Score	0.692	0.930	0.817	0.705	0.646	0.364
	Balanced Accuracy	0.870	0.880	0.969	0.804	0.944	0.755
	Kappa	0.694	0.709	0.799	0.677	0.634	0.349
Hybrid-XGB, Base	Accuracy	0.843	0.877	0.954	0.923	0.966	0.966
	Sensitivity	0.721	0.897	0.942	0.386	0.913	0.467
	Specificity	0.937	0.805	0.955	0.981	0.968	0.975
	F1-Score	0.615	0.919	0.760	0.496	0.583	0.318
	Balanced Accuracy	0.829	0.851	0.949	0.684	0.940	0.721
	Kappa	0.611	0.661	0.736	0.458	0.568	0.303

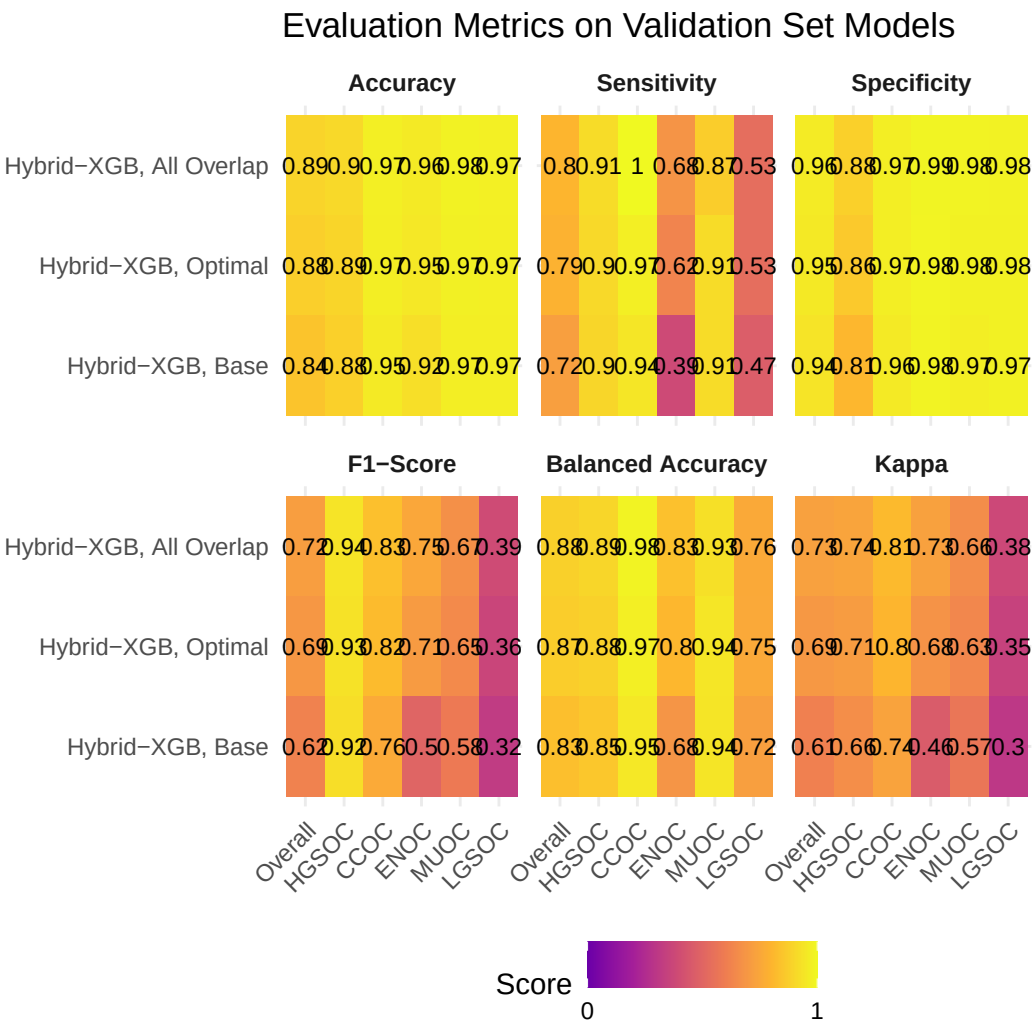


Figure 4.21: Evaluation Metrics on Validation Set Models

4.5.2 Confusion Matrices

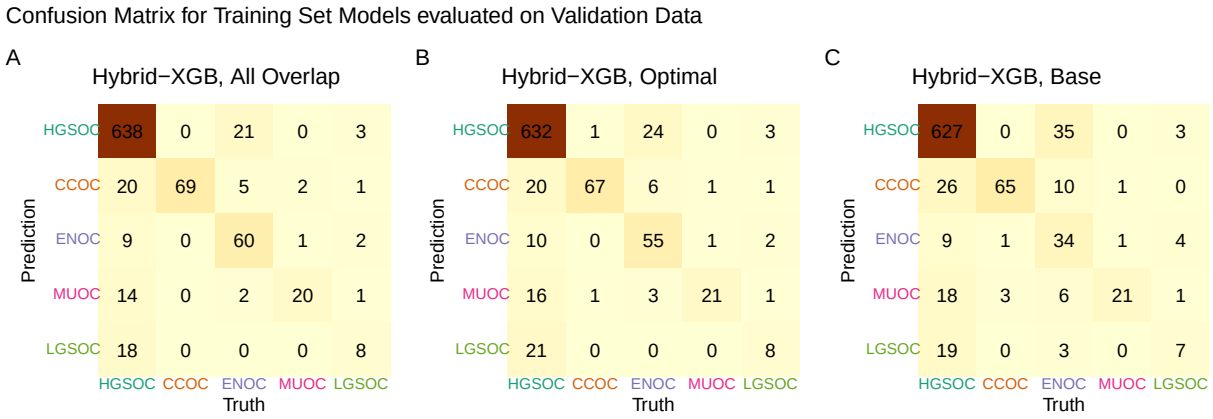


Figure 4.22: Confusion Matrix for Training Set Models evaluated on Validation Data

4.5.3 ROC Curves

4.5.3.1 All Overlap

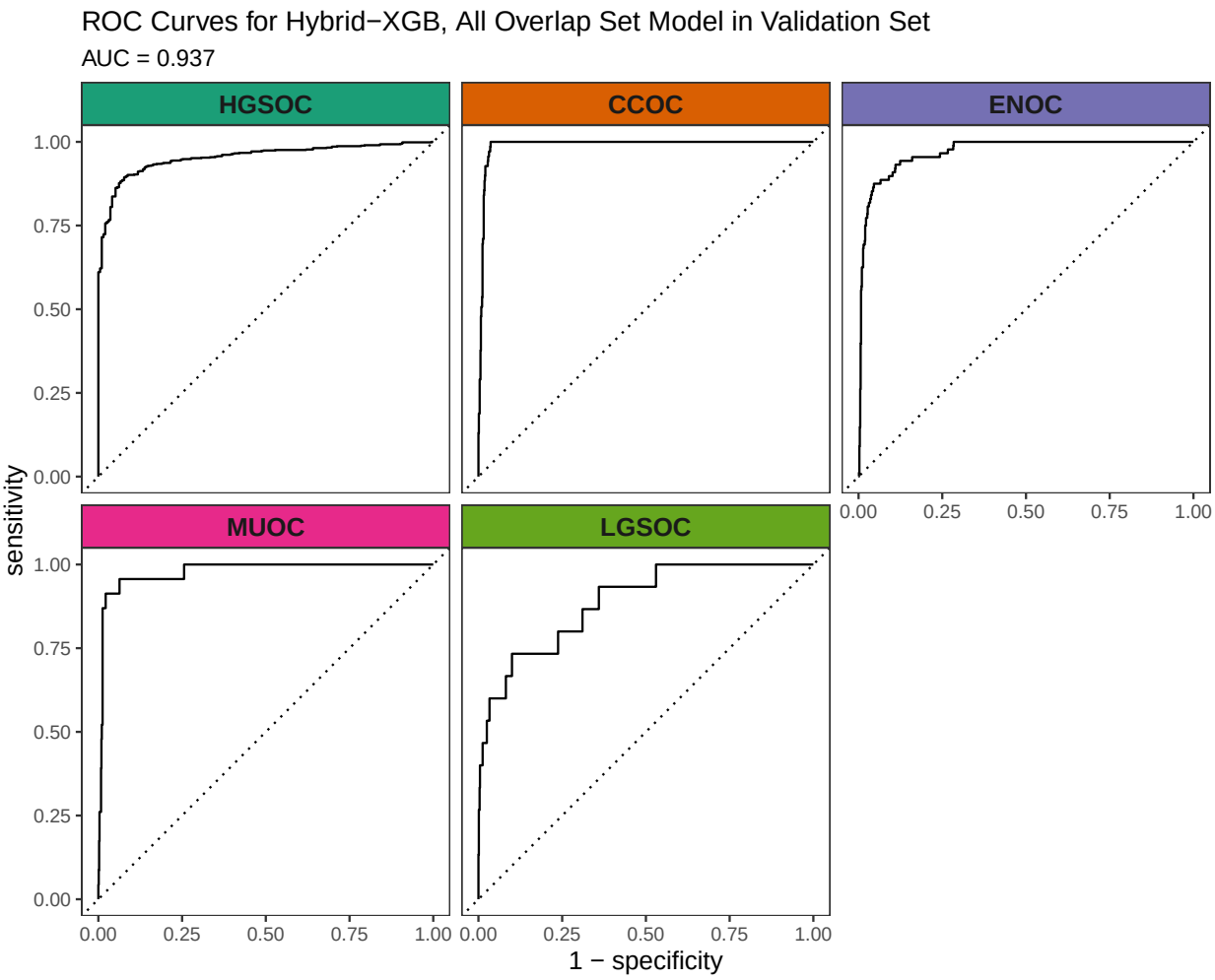


Figure 4.23: ROC Curves for Hybrid-XGB, All Overlap Set Model in Validation Set

4.5.3.2 Optimal Set

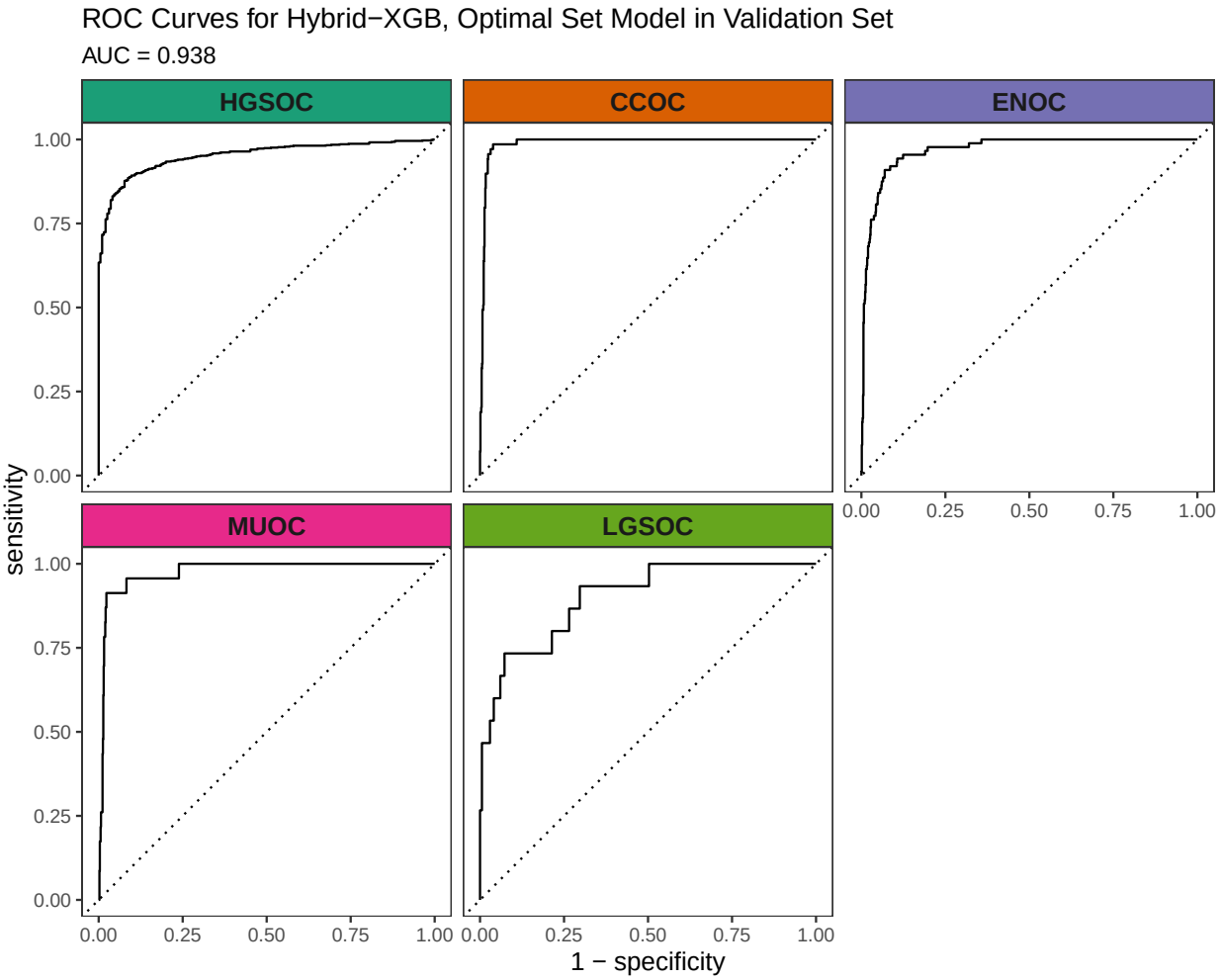


Figure 4.24: ROC Curves for Hybrid-XGB, Optimal Set Model in Validation Set

4.5.3.3 Base Set

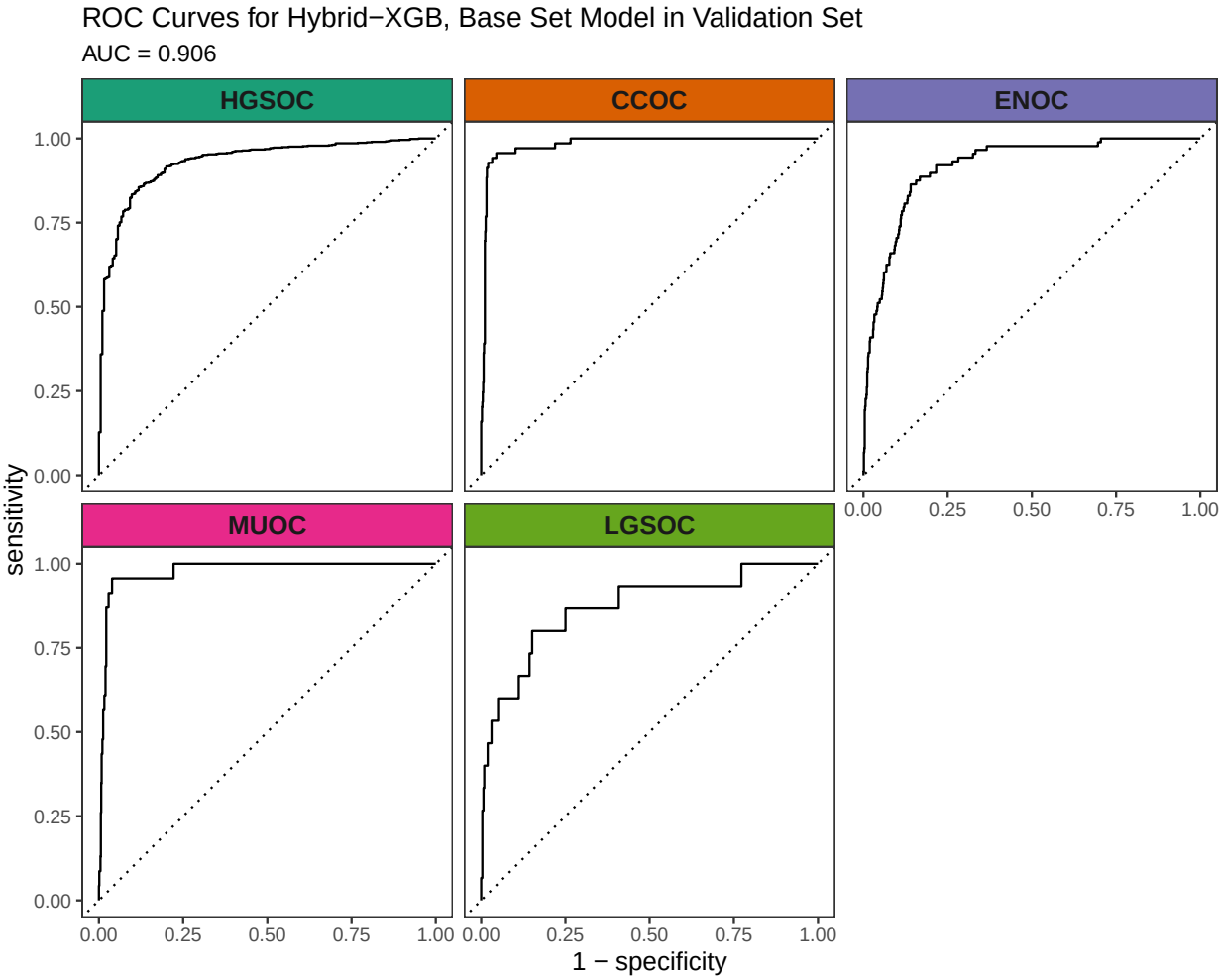


Figure 4.25: ROC Curves for Hybrid-XGB, Base Set Model in Validation Set

4.5.4 Calibration Plots

4.5.4.1 All Overlap

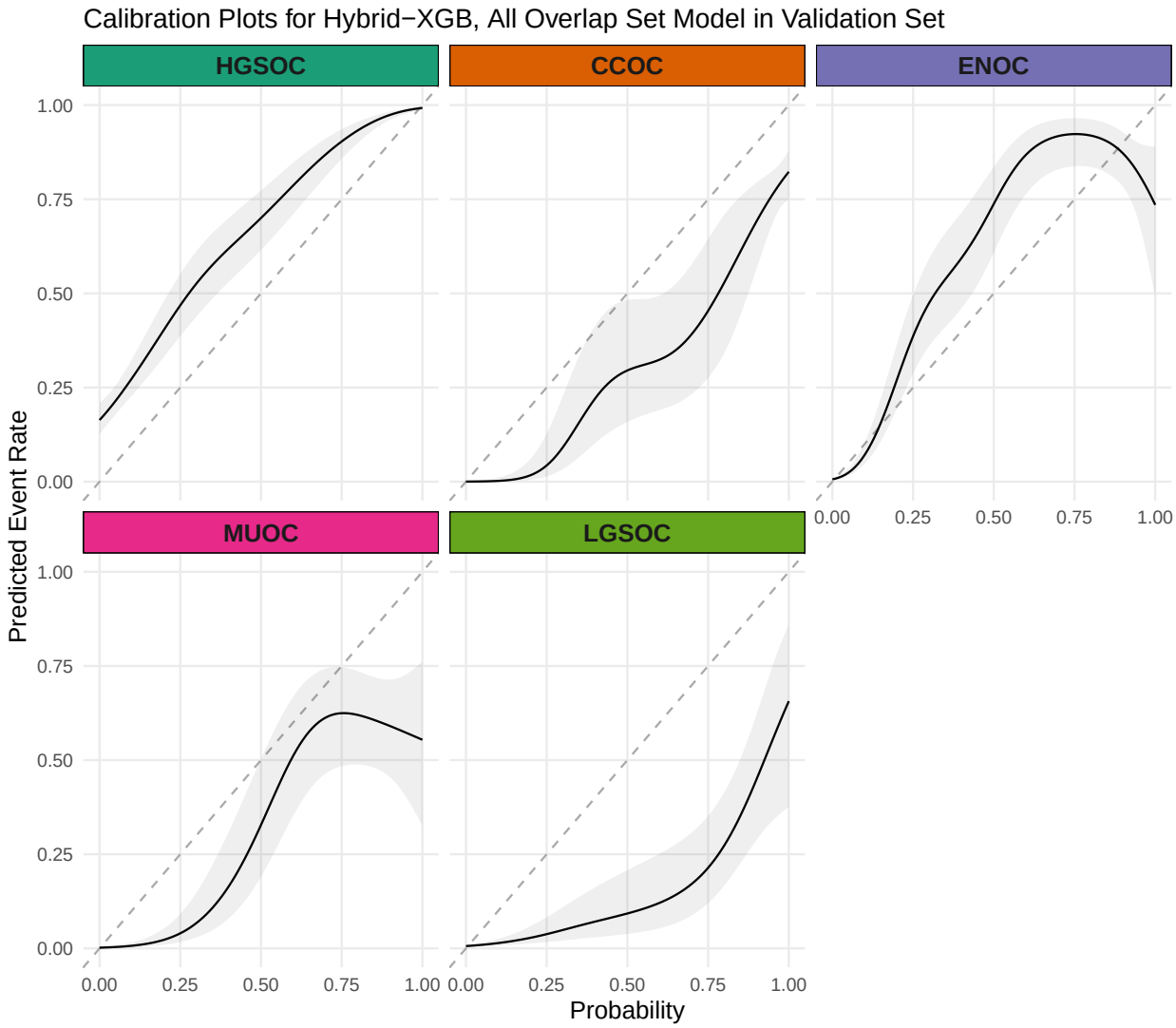


Figure 4.26: Calibration Plots for Hybrid-XGB, All Overlap Set Model in Validation Set

4.5.4.2 Optimal Set

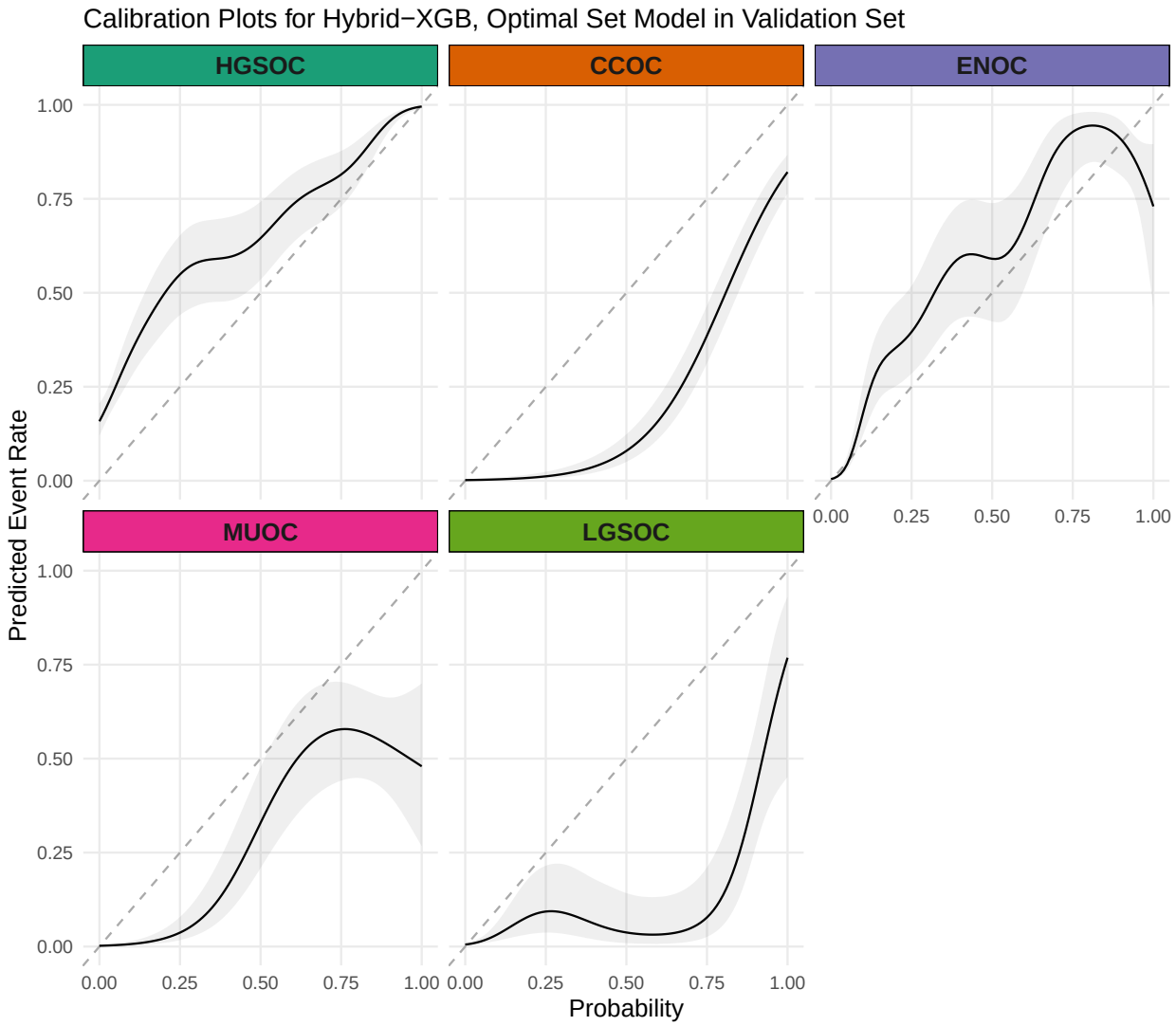


Figure 4.27: Calibration Plots for Hybrid-XGB, Optimal Set Model in Validation Set

4.5.4.3 Base Set

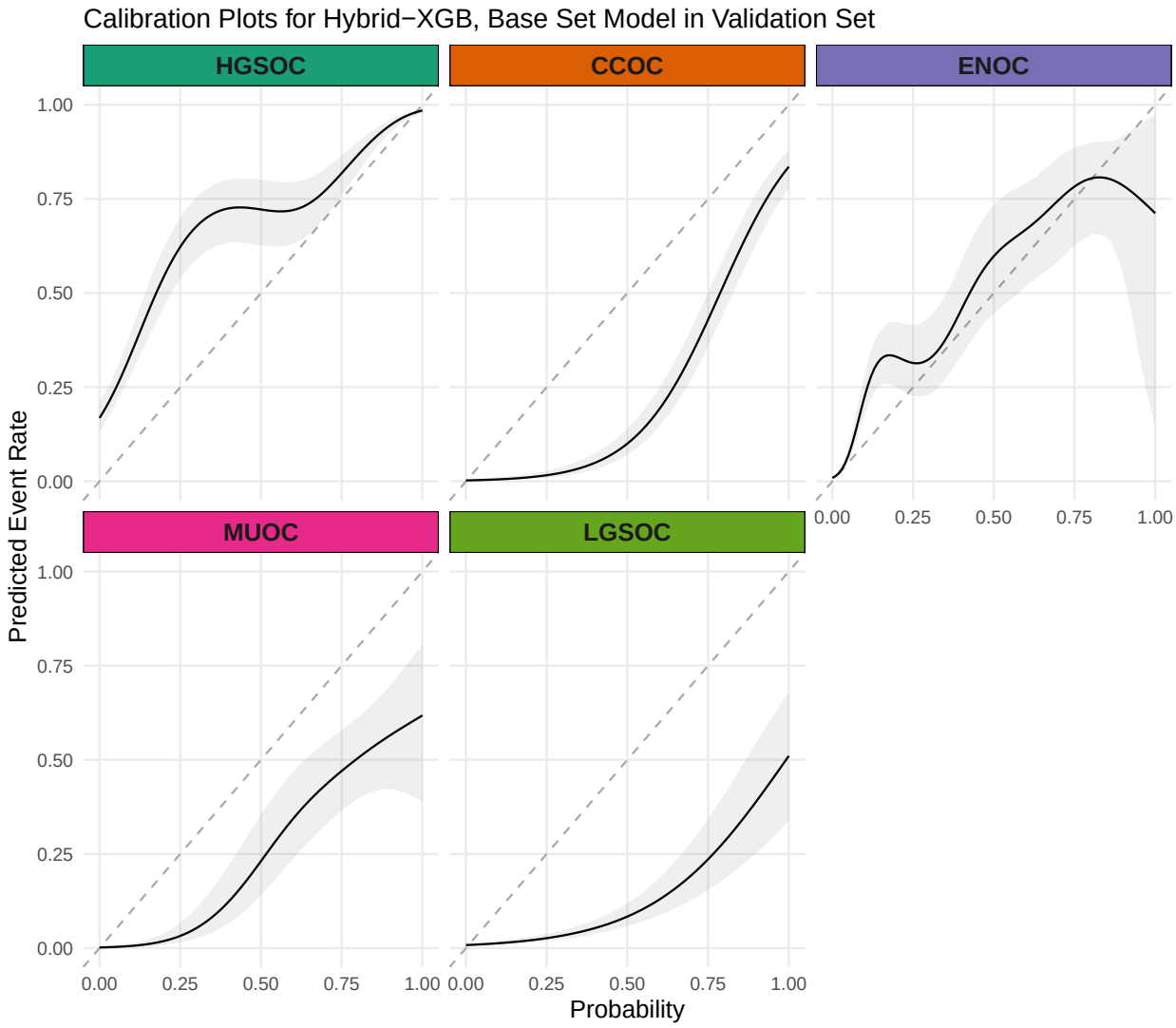
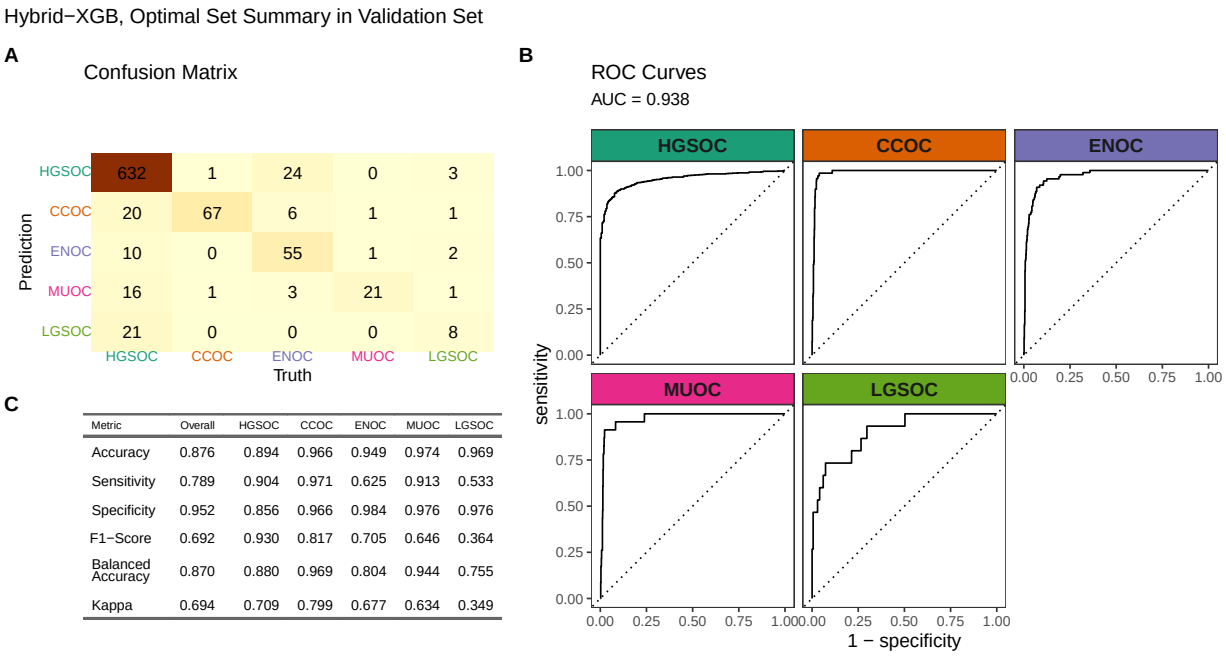


Figure 4.28: Calibration Plots for Hybrid-XGB, Base Set Model in Validation Set

4.5.5 Summary

A summary of the Hybrid-XGB, Optimal model results are shown in Figure 4.29.



C

Metric	Overall	HGSOC	CCOC	ENOC	MUOC	LGSOC
Accuracy	0.876	0.894	0.966	0.949	0.974	0.969
Sensitivity	0.789	0.904	0.971	0.625	0.913	0.533
Specificity	0.952	0.856	0.966	0.984	0.976	0.976
F1−Score	0.692	0.930	0.817	0.705	0.646	0.364
Balanced Accuracy	0.870	0.880	0.969	0.804	0.944	0.755
Kappa	0.694	0.709	0.799	0.677	0.634	0.349

Figure 4.29: Validation Summary

4.5.6 Additional Explorations

Table 4.19: Clinicopath characteristics between correct and incorrect predictions of ENOC cases

Characteristic	**Predicted ENOC Correctly** N = 55[1]	**Missed ENOC** N = 33[1]	**p**
Age at diagnosis	53 (46, 63)	56 (51, 62)	
Tumour grade			
low grade	42 (93%)	17 (63%)	
high grade	3 (6.7%)	10 (37%)	
Unknown	10	6	
FIGO tumour stage			
I	42 (78%)	20 (61%)	
II-IV	12 (22%)	13 (39%)	
Unknown	1	0	
Race			
white	50 (93%)	26 (90%)	
non-white	4 (7.4%)	3 (10%)	
Unknown	1	4	
ARID1A			
absent/subclonal	10 (18%)	6 (18%)	
present	45 (82%)	27 (82%)	
WT1			
diffuse (>50%)	2 (3.6%)	3 (9.1%)	
focal (1-50%)	3 (5.5%)	0 (0%)	
negative	50 (91%)	30 (91%)	
TP53			
mutated	1 (1.8%)	4 (13%)	
wild type	54 (98%)	28 (88%)	
Unknown	0	1	
PR			
diffuse (>50%)	35 (64%)	11 (33%)	
focal (1-50%)	9 (16%)	6 (18%)	
negative	11 (20%)	16 (48%)	
P16			
abnormal block	2 (3.6%)	4 (12%)	
abnormal complete absence	13 (24%)	12 (36%)	
normal	40 (73%)	17 (52%)	
NAPSIN A			
negative	53 (96%)	32 (100%)	
positive	2 (3.6%)	0 (0%)	
Unknown	0	1	

[1]Median (Q1, Q3); n (%)
[2]Wilcoxon rank sum test; Fisher's exact test; Pearson's Chi-squared test

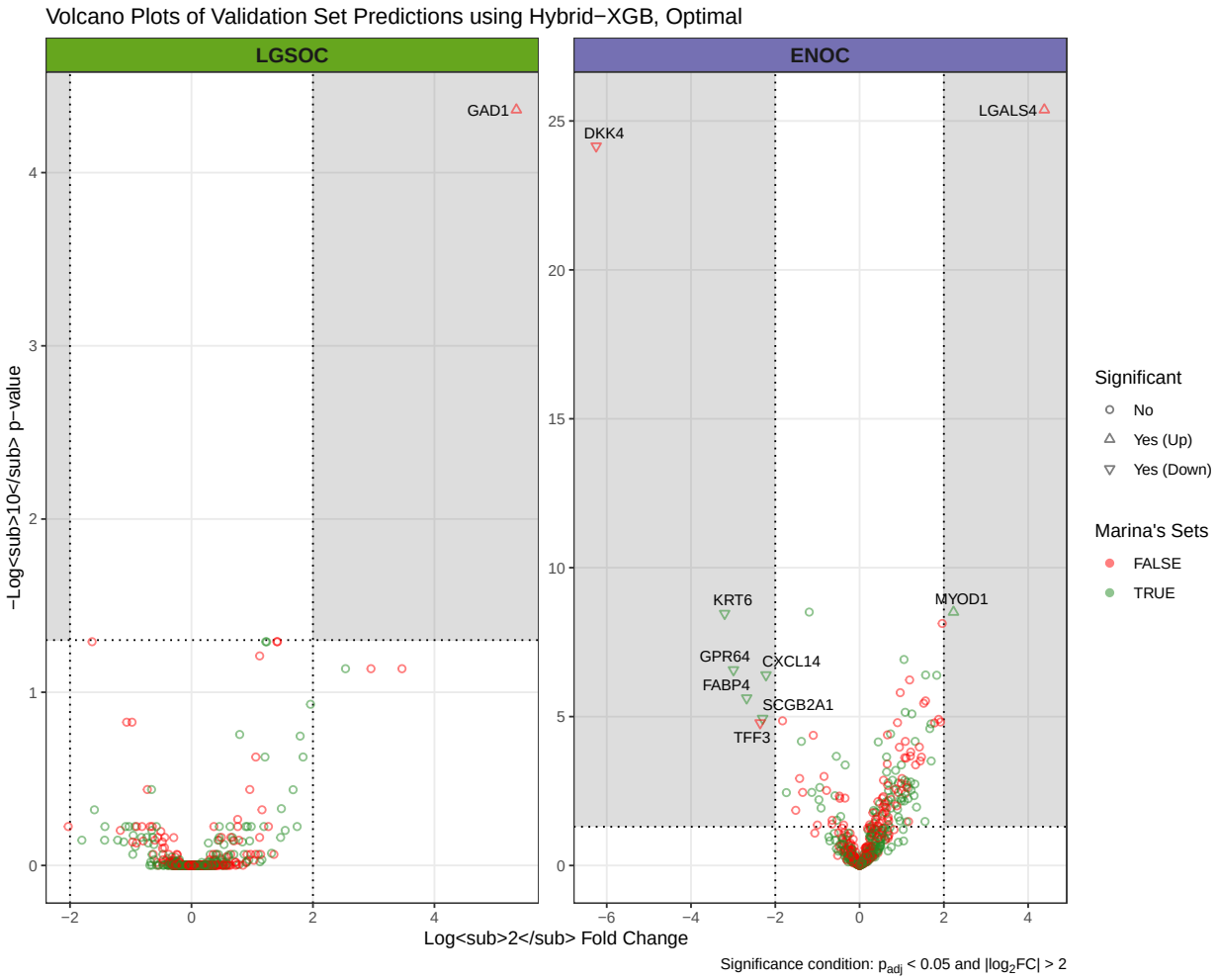


Figure 4.30: Volcano Plots of Validation Set Predictions

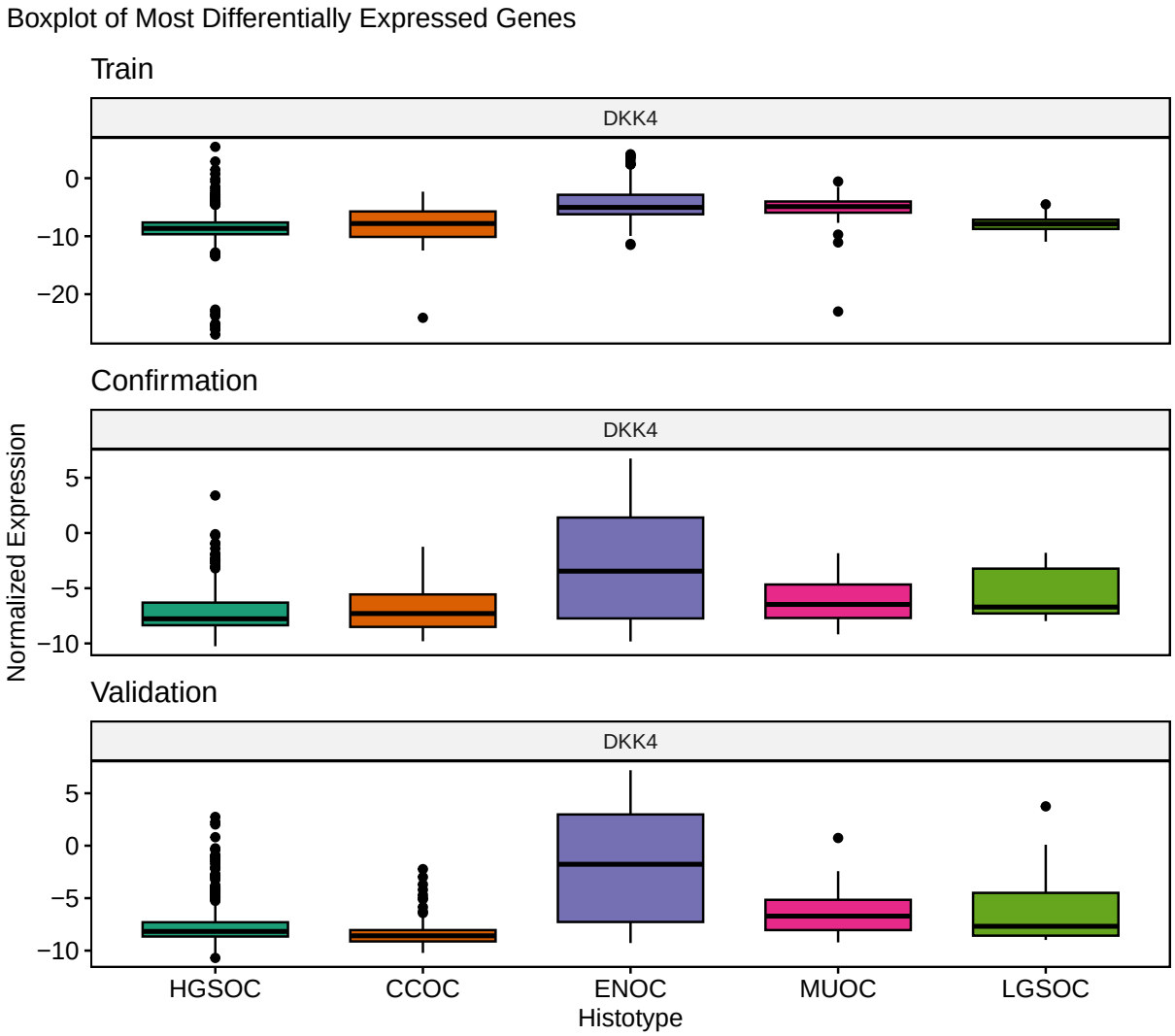


Figure 4.31: Boxplot of Most Differentially Expressed Genes

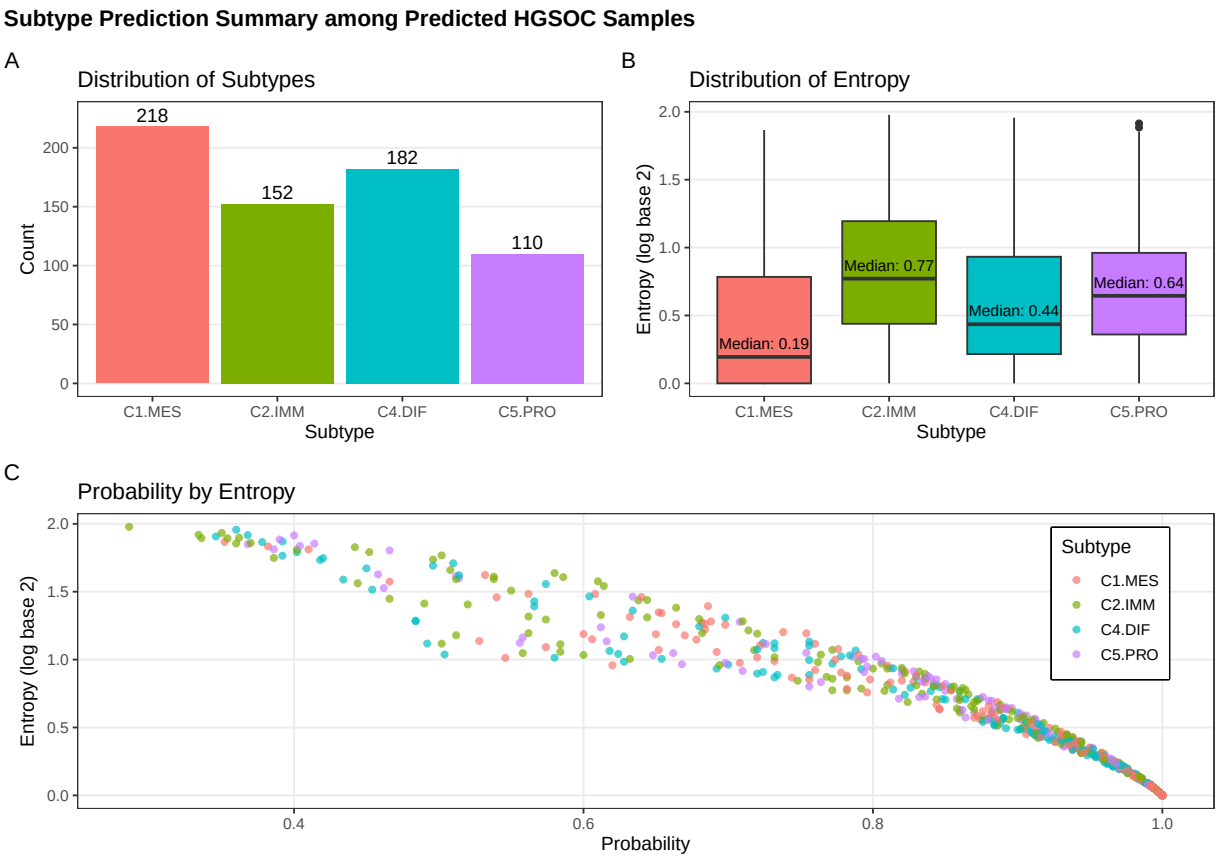


Figure 4.32: Subtype Prediction Summary among Predicted HGSOC Samples

References

Pihur, Vasyl, Susmita Datta, and Somnath Datta. 2009. "RankAggreg, an r Package for Weighted Rank Aggregation." *BMC Bioinformatics* 10 (1): 62.

Talhouk, Aline, Stefan Kommoss, Robertson Mackenzie, et al. 2016. "Single-Patient Molecular Testing with NanoString nCounter Data Using a Reference-Based Strategy for Batch Effect Correction." *PLOS ONE* 11 (4): e0153844. https://doi.org/10.1371/journal.pone.0153844.